치의과학석사 학위논문

항암 치료 반응과 유전자 변이 조합의 상관성 분석 도구 구현

An analysis tool of co-occurring genetic alteration driven clinical response of anti cancer therapy

2020년 8월

서울대학교 대학원

치의과학과 의료경영과정보학전공

손 지 원

국 문 초 록

항암 치료 반응과 유전자 변이 조합의 상관성 분석 도구 구현

손지원
서울대학교 의료경영과정보학전공

서론: 암세포의 유전체 정보를 통해 개인에게 가장 적합한 항암 치료 방법을 추천하는 것을 정밀 종양학(precision oncology)이라고 한다. 유전자 변이 정보는 해당 암 조직의 세포학적 특성을 파악을 가능하게 하여 개인의 항암 치료 반응을 예측할 수 있도록 돕는다. 이처럼 정밀 종양학의 주요한 생물지표(biomarker)인 유전자 변이 정보는 개인 맞춤형 치료를 가능하게 해주었으나, 한계 또한 존재한다. 같은 유전자 변이를 기준으로 분류 및 투약된 환자들도 치료 반응이 상이한 경우가 다수 관찰된다. 이는 암의 이종성(cancer heterogeneity) 때문이며, 단일 유전자 변이가 각 환자의 치료 반응을 예측하는 생물지표로써 한계가 있음을 의미한다. 따라서 본 연구는 단일 이상의 다수의 유전자 변이 조합이 항암 치료 반응을 구분하는 생물지표로서의 활용 가능성을 확인하고자 하였다.

방법: 본 연구는 임상 데이터 획득, 데이터 전처리, 유전자 변이 조합 추출과 치료 반응 맵핑, 조합과 치료 반응의 유의도 계산 및 랭킹 목록 추출 과정으로 진행된다. 임상 데이터의 유전자 변이 데이터와 환자/검체 메타 데이터를 획득하고, 전처리 과정을 거쳐 빈발 아이템 추출(Frequent Itemset Mining) 방법을 활용하여 조합을 추출하였다. 기준 지지도(support)값 이상으로 발생한 조합을 각 환자의 유전자 변이 데이터와 맵핑하여 해당 조합을 가진 환자들의 목록을 추출하였다. 치료 반응 데이터와 환자 데이터를 대응하여 조합별 치료 반응이 있는 그룹, 치료 반응이 없는 그룹의 환자 수를 셈하였다. 유전자 변이 조합과 치료

반응 간의 상관성을 계산하기 위해 방법론으로 베이즈 요인(Bayes Factor)을 사용하여 값을 도출하였다. 최종적으로 항암 치료 반응과 상관성이 있는 순으로 변이 양상 목록을 정렬하여 제공한다. CBioPortal에서 제공하는 공개 데이터를 사용하였으며, 그 종류는 표적 치료 임상 데이터 2개, 면역 치료 임상 데이터 2개, 간 절제술 후 재발 여부 데이터 1개이다.

결과: 임상 데이터를 기반으로 유전자 변이 조합과 치료 반응의 상관성을 계산하여 제공하는 도구를 구현하였다. 각 데이터를 분석한 결과 항암 치료 반응과 상관성이 있는 유전자 변이 양상을 단일부터 조합까지 다양하게 추출할 수 있었다. 추출된 유전자 변이 조합은 문헌 검색을 통해 그 유의성을 검증하였다. 또한, 단일로는 상관성을 보이지 않았던 유전자 변이가 다른 유전자 변이와의 조합을 이룰 때 높은 상관성을 보이는 현상을 확인할 수 있었다. 이는 항암 치료 반응의 생물지표로써 단일을 넘어 그 이상의 유전자 변이 조합의 활용 가능성과 필요성을 보여주는 결과라고 할 수 있다.

주요어 : 정밀 종양학, 유전자 변이, 항암 치료, 치료 반응,
빈발 항목 집합 추출, 암 유전체학

학 번 : 2016-22035

목 차

표 목록

그림 목록

서 론

1. 연구의 필요성

정밀 의학(precision medicine)이란 환자 개개인의 유전체 정보, 병력, 생활 습관, 환경에 맞추어 질병의 예방과 치료를 하는 것을 말한다. 정밀 의학은 차세대 염기서열 분석(NGS, Next Generation Sequencing)과 같은 유전체 염기서열 분석 방법과 전자 의무 기록(EMR, Electronic Medical Record), 그리고 빅데이터의 발달과 그 궤를 같이한다. 환자의 메타데이터, 유전체 정보 데이터를 대용량으로 수용할 수 있게 되고, 이에 다양한 분석 방법론이 활용됨으로써 개인 맞춤 의료의 시대가 열리게 된 것이다[1].

정밀 종양학(precision oncology)은 정밀 의학을 암에 적용한 것으로, 특히 환자 암세포의 유전체 분석 결과, 암세포의 메타데이터를 중심으로 치료 및 예방 방법을 달리하는 것을 말한다[2]. 정밀 종양학의 발달은 미국 국립 암 센터(NCI, National Cancer Institute)가 주관한 TCGA(The Cancer Genome Atlas)[3] 프로젝트와 같이 국가적으로 진행된 암 환자 유전체 데이터베이스의 공개와 그 활용으로 인해 가속화되었다. 암 유전체 데이터는 유전자 돌연변이(gene mutation) 데이터, 유전자 발현 데이터(gene expression), 데이터, DNA 메틸레이션(DNA methylation) 데이터, CNA(Copy Number Alteration) 등으로 나뉘며 데이터별로 그 활용 방안이 다르다. 이 중 유전자 돌연변이란 암 조직 혹은 혈액순환종양 DNA 등에서 추출된 암세포의 유전자의 돌연변이를 말한다. 변이된 유전자로부터 발현된 단백질은 정상 세포와는 다른 메커니즘을 유도하여 암의 발생, 유지, 전이 등에 영향을 미칠 수 있다. 유전자 변이 데이터는 이러한 단백질 변이를 파악할 수 있는 기반이 될 수 있기에 중요한 생물지표(biomarker)로 꼽힌다[4].

유전자 변이는 항암 치료의 치료 반응을 예측하고 환자에게 적합한 치료법을 추천하는 데 있어 주요한 역할을 한다. 유전자 변이 양상을 통해

암세포의 세포학적 특성을 파악할 수 있기 때문이다[4]. 암을 유발하는 유전자(oncogene)가 과발현된 암세포와 암을 억제하는 유전자(tumor suppressor)가 저해된 암세포는 그 치료법이 다르다.

항암 요법 중 표적 치료 요법은 항암제가 변이된 단백질에 직접 작용하도록 하는 것으로, 표적 단백질을 탐색하는 과정부터 치료제를 설계하는 과정에 단백질 변이 및 유전자 변이를 파악하는 것이 필수적이다. 환자에게 실제 투약을 권할 때도 해당 항암제가 표적으로 하는 유전자 변이의 여부를 기준으로 하게 된다. 면역 치료 중 면역 체크포인트 치료(immune checkpoint therapy)의 경우 암세포의 유전자 변이 정도를 정량적으로 계산한 TMB(tumor mutation burden)값과 치료 반응이 관계가 있는 것으로 알려져 있으며[5], 이는 암세포가 TMB값을 높일 수 있는 유전자 변이가 얼마나 있는지에 따라 예측될 수 있다[6]. 따라서 높은 TMB와 관련이 있는 유전자 변이를 가진 환자에게 면역 체크포인트 치료법을 권할 수 있다.

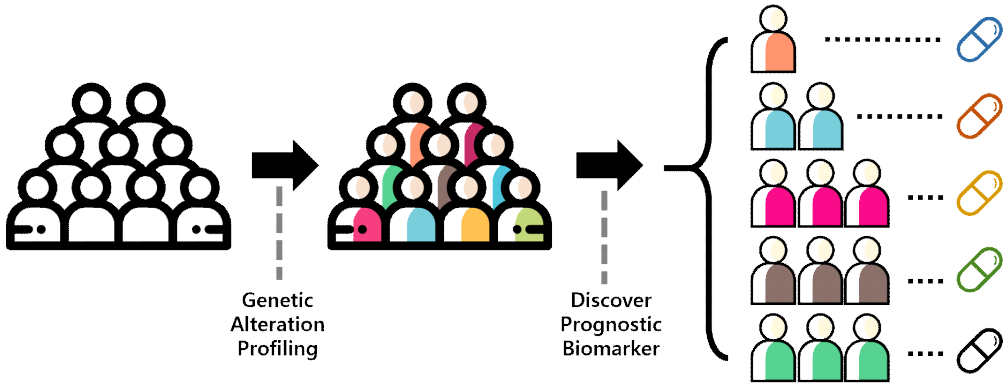
이처럼 항암 치료 요법과 유전자 변이 정보의 관계성을 밝히려는 연구는 계속되고 있다. 이때 유전자 변이란 주로 단일을 기준으로 해왔으나, 단일 변이를 기준으로 투약 시 그 치료 효과가 개인마다 상이하며 치료 반응을 예측할 수 없다는 한계점이 있었다. 이를 극복하기 위해 생물학적 패스웨이와 단백질 간 상호작용 연구 등 다수의 유전자 변이로 인한 생물학적 파급효과를 연구하는 분야가 확장되었다. 다중 생물지표(multi biomarker)로서 유전자 변이를 다루고자 하는 시도가 증가하게 된 것이다.

Mina, M et el.의 연구 SELECT(Selected Events Linked By Evolutionary Conditions across human Tumors)[7]는 암의 유전자 변이 양상을 단일 유전자가 아닌 유전자 쌍(pair)과 항암 치료제 간의 상관성을 증명한 연구이다. 암세포에서 함께 자주 관측되는 유전자 변이(co-mutation)가 우연에 의한 사건이 아닌 유의미한 패턴을 형성하며, 이는 암의 진화적 의존(cancer evolutionary dependency)에 의한 결과라는 것이 해당 연구의 골자이다. 암의 진화적 의존이란 암이 형성되고 성장하는 과정에서 해당 암세포가 가진 유전자 변이와 암 미세환경(tumor

microenvironment), 암의 면역 반응(immune response) 등이 추후 유전자 변이 양상에 영향을 주는 것을 생물학적 진화의 관점에서 해석하는 것을 말한다[8]. 해당 연구는 대용량 암 데이터에서 추출한 유전자 변이 쌍이 암의 진화적 의존에 의해 선택된 쌍일 경우, 항암 치료 반응이 좋지 않음을 세포주 실험을 통해 밝혀냈다[7].

본 연구는 다중 유전자 변이와 항암 치료 반응과의 상관성을 실제 임상 데이터를 바탕으로 분석할 수 있는 도구를 구현하고자 하였다. 임상 데이터의 환자별 유전자 변이 데이터와 치료 반응데이터를 활용해 함께 자주 발생하는 유전자 변이(co-mutation)의 조합을 추출하고, 이 조합을 가진 환자들의 치료 반응데이터를 이와 맵핑하여 상관성을 계산하였다. 적은 데이터에서 데이터의 가설 지지도 및 편향성을 수치화하여 파악할 수 있는 베이즈 요인(bayes factor)을 상관성 계산에 사용하였다. 최종적으로 유전자 변이 조합과 치료 반응 유무의 상관성 값을 기준으로 정렬된 리스트(ranked list)로 제공한다. 추출된 결과는 해당 임상 데이터의 항암 치료법이 특정 환자에게 치료 반응이 있을지를 특정 유전자 변이 조합의 유무를 통해 구분할 수 있도록 돕는다. 추후 본 도구를 활용하여 다양한 임상 데이터 분석 결과가 누적된 대규모 항암 치료 반응 데이터베이스를 구축할 수 있을 것이다. 이를 통해 개인의 유전자 변이 양상에 맞추어 특정 유전자 변이 조합의 유무에 따라 가장 적합한 항암 치료 요법을 추천해 줄 수 있을 것으로 기대한다.

그림 1. 정밀 종양학에서 유전자 변이 정보의 활용



2. Frequent Itemset Mining

Frequent Itemset Mining(빈발 항목 집합 발굴)은 Frequent Pattern Mining(빈발 패턴 발굴) 방법의 하나로, 대용량 데이터에서 함께 자주 발생하는 항목(item)을 찾아내 그 패턴을 발굴하는 데 사용된다[9]. 대형 슈퍼마켓에서 고객들의 거래 내역의 데이터를 저장하였을 때, 고객들이 구매한 상품이 항목이며, 한 거래 건을 트랜잭션(transaction)이라고 한다. 표 1과 같이 각 거래에 대한 고유한 ID를 부여하고 고객이 구매한 상품의 목록을 나열한 형태의 데이터가 있다고 가정한다. 해당 데이터를 기준으로 빈발(다빈도) 항목 발굴을 하게 되면, 고객들이 자주 함께 구매하는 상품의 패턴을 알 수 있다.

데이터에서 다빈도 항목을 추출할 때는 최소 발생 빈도를 설정해야 한다. 최소 발생 빈도를 지지도(support)라고 하며, 지지도를 기준으로 유의한 항목을 추출하게 된다. 표 1 데이터의 경우, 지지도를 0.6이라고 설정했을 때 추출되는 항목의 결과는 표 2와 같다. 5개의 거래중 3건 이상에 포함되어있는 {A}, {C}, {E}, {A, C}, {A, E}, {C, E}, {A, C, E}이 추출된다.

표 1. 마트 구매 이력 트랜잭션 예시

트랜잭션 ID (거래 ID)	항목 목록 (거래 내용)
1	A, B, C, D, E, F
2	A, C, E, G
3	A, E, G, H
4	A, C, E, I
5	B, H

표 2. 지지도를 기준으로 추출된 항목 목록

다빈도 항목	지지도 (support)
{A}	4 / 5 = 0.8
{C}	3 / 5 = 0.6
{E}	4 / 5 = 0.8
{A, C}	3 / 5 = 0.6
{A, E}	3 / 5 = 0.8
{C, E}	3 / 5 = 0.6
{A, C, E}	3 / 5 = 0.6

본 연구에서는 Frequent Itemset Mining 방법을 활용하여 항암 치료제 임상 데이터에서 환자 유전자 변이 검사 데이터를 기반으로 다빈도 항목 집합을 추출하였다. 이때 트랜잭션 ID는 환자 ID가 되고, 변이 유전자는 항목(item)이 된다. 추출된 유전자 변이 조합과 해당 조합을 가진 환자들의 치료 반응데이터를 기반으로, 베이즈 요인 계산법을 활용하여 유전자 변이 조합과 치료 반응 간의 상관성을 발굴하고자 하였다.

연구 방법

그림 2. 연구 방법 순서도

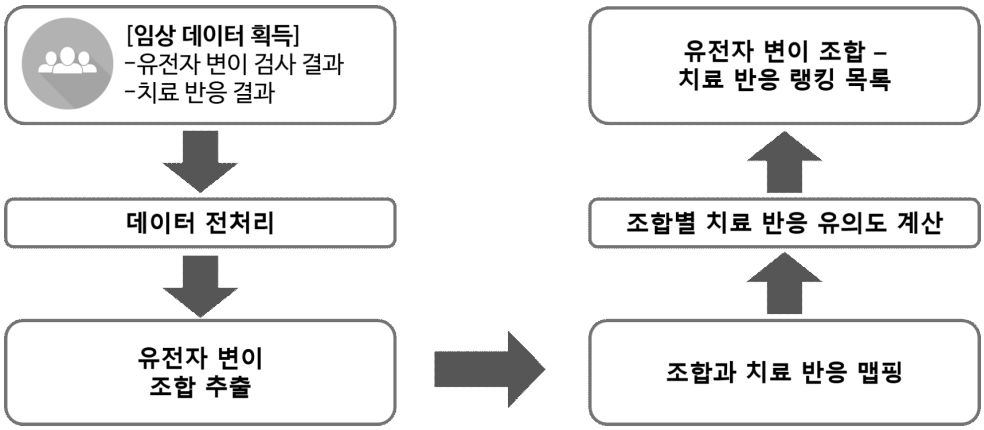


그림 2는 연구방법의 순서를 모식도로 표현한 것이다. 본 연구는 (1) 임상 데이터 획득 (2) 데이터 전처리 (3) 유전자 변이 조합 추출 (4) 조합과 치료 반응 맵핑 (5) 조합별 치료 반응 유의도 계산 (6) 유전자 변이 조합 - 치료 반응 랭킹 목록 추출의 순서로 이루어진다. 모든 과정은 JAVA로 작성되었으며, 모듈화되어 실행된다.

1. 임상 데이터 획득

본 연구는 유전자 변이 조합과 다양한 항암 요법의 치료 반응의 상관성 추출 결과를 보여주기 위해 총 5개의 임상 데이터를 선택하였다. 그 종류와 수는 표적 치료 임상 데이터 2개, 면역 체크포인트 치료 임상 데이터 2개, 수술 요법 임상 데이터 1개이다. 표적 치료 요법은 변이된 단백질을 표적으로 하여 치료하는 방법으로, 추출 결과를 표적 단백질을 중심으로 해석할 수 있다. 면역 치료 요법은 치료 반응과 관계가 있다고 알려진 각종 생물지표(TMB 등)와 추출된 유의한 유전자 변이 조합의 관계를 해석할 수 있다[5]. 수술 요법은 앞선 두 가지의 항암 요법과는

다른 물리적인 접근 방법이나, 수술 후 재발 여부 등을 예측하는데 암세포의 유전자 변이 양상을 활용할 수 있다.

데이터는 암 유전체학 연구를 위한 임상 데이터를 배포하는 CBioPortal[11]에서 획득하였다. CBioPortal은 메모리얼 슬론 케터링 암센터(MSK, Memorial Sloan Kettering Cancer Center)에서 개발한 암 임상 데이터 플랫폼이다. 충분한 유전자 변이 조합을 추출하기 위해 전체 환자 수가 50명 이상이며, 환자 대부분의 유전자 변이 검사 결과와 치료 반응데이터가 공개된 데이터를 선택하였다. 또한, 치료 반응 평가 지표로 항암 요법의 치료 반응을 판단하는 지표인 RECIST(Response Evalution Criteria In Solid Tumors)[11] 혹은 치료 반응이 있음/없음으로 이진법적(binary)으로 구분될 수 있는 지표를 사용한 데이터를 활용하였다.

가. 표적 치료 요법 임상 데이터

본 연구에서 사용된 표적 치료 요법 임상 데이터는 총 2가지이다. 데이터①은 유전자 ERBB2 혹은 ERBB3 변이를 가진 환자 141명을 대상으로 ERBB protein family kinase domain 억제제인 neratinib을 투여한 바스켓 임상시험(basket trial) 이다[12]. 바스켓 임상시험이란, 암종과 관계없이 같은 유형의 유전자 변이를 가진 환자들에게 동일한 항암제를 투약하는 임상시험 방식을 말한다.

데이터②는 호르몬 수용체 양성(HR+)인 국소 진행성 유방암 혹은 전이 유방암 환자 51명을 대상으로 PIK3CA 억제제인 alpelisib과 항 에스트로겐 치료제인 아로마타아제 억제제를 함께 투여한 결과이다[13].

나. 면역 체크포인트 치료 요법 임상 데이터

면역 체크포인트 치료 요법이란 항원 특이적 면역 세포인 T 림프구(T-cell)가 암에 정상적으로 작용하여 제거할 수 있도록 돕는 것이다. T 림프구의 PD-1 혹은 CTLA-4은 정상 세포 여부를 판단하는 면역관문(checkpoint) 역할을 한다. PD-1 단백질은 정상 세포의 PD-L1 단백질과 결합하며, CTLA-4은 B7-1/B7-2 단백질과 결합하게 된다. 암세포 중 일부는 정상 세포가 가진 PD-L1 혹은 B7-1/B7-2 단백질이 있어 T 림프구의 면역 반응을 피해 체내 면역 감시망을 피해갈 수 있다. 면역 체크포인트 치료 요법은 환자에게 PD-1, PD-L1, CTLA-4 억제제를 투여하여 암세포가 정상 세포로 위장하는 것을 막아 결과적으로 암세포가 체내 면역 체계에 의해 저해되도록 한다[14].

본 연구에서 사용된 면역 체크포인트 치료 임상 데이터①은 249명의 환자를 대상으로 PD-1, PD-L1, CTLA-4 억제제를 사용하였다[15]. 환자의 암종 분포는 악성 흑색종 151명, 폐암 57명, 방광암 27명, 두경부 편평세포암종 12명, 기타 암종 2명으로 구성되었다. 데이터②는 비소세포 폐암 환자 240명을 대상으로 PD-1, PD-L1 억제제를 투여한 결과이다[16].

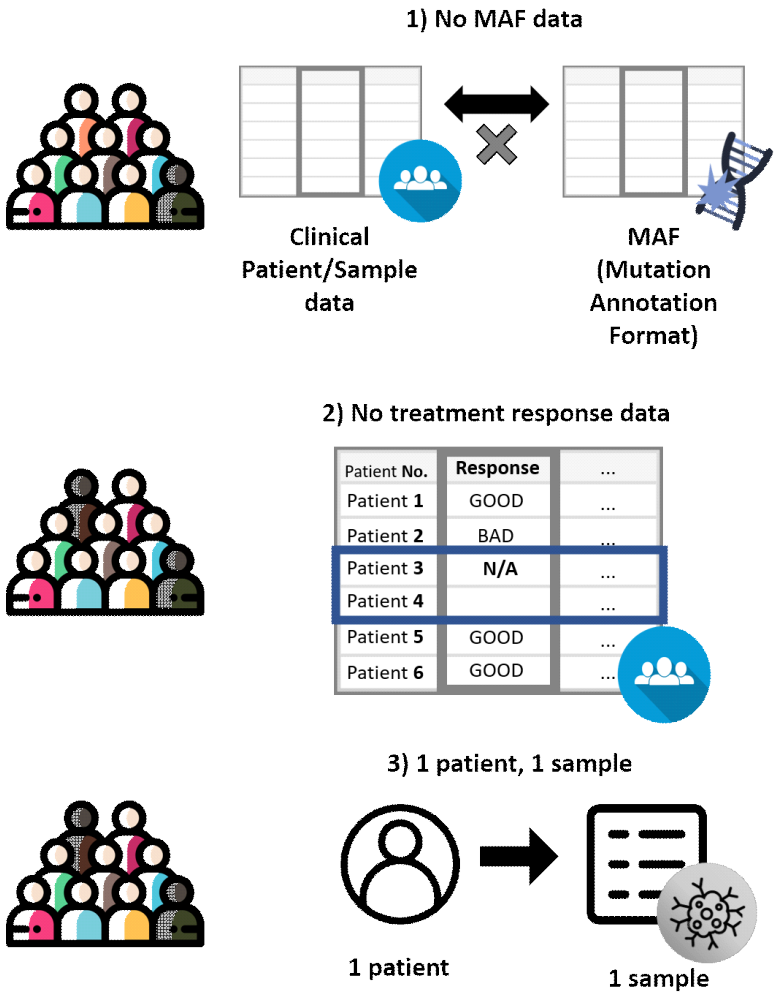
다. 간 절제술 후 재발 여부 데이터

해당 데이터는 236명의 초기 간세포 암(early stage hepatocellular carcinomas) 환자를 대상으로 간 절제술(hepatic resection) 후 재발 여부를 관측한 데이터이다. 간 절제술은 초기 간세포 암 환자에게 가장 효과적인 치료법으로 알려졌지만, 수술 후 5년 이내 50% 이상의 높은 재발률을 보인다. 해당 데이터는 간 절제술 후 재발을 유전자 변이 양상을 통해 예측하고자 위한 목적으로 수집되었다[17].

2. 데이터 전처리

가. 데이터 필터링

그림 3. 데이터 필터링 조건



1) 유전자 변이 검사 데이터가 없는 환자 제외

환자 혹은 검체 메타데이터와 치료 반응데이터가 존재하더라도 해당 환자의 유전자 변이 검사 데이터가 존재하지 않을 수 있다. 환자/검체 메타데이터의 환자/검체 ID를 MAF 데이터, CNA 데이터와 대조하여 유전자 변이 검사 데이터가 없는 환자일 때 분석 대상에서 제외한다. 해당 작업을 거치지 않으면 유전자 변이 조합과 치료 반응데이터의 상관성 계산 시 총환자 수 계산에 오류가 생길 수 있다.

2) 치료 반응데이터 값이 없는 환자 제외

본 연구는 치료 후 반응데이터를 활용하였기 때문에, 환자 및 검체 데이터에 치료 반응데이터값이 기록되어있지 않거나, 'NA(값 없음)'일 경우 해당 환자를 포함하지 않았다. 치료 반응데이터가 없는 환자의 유전자 변이 검사 결과를 포함하게 되면, 상관성을 계산할 수 없는 조합의 빈도수가 증가하기 때문이다. 이는 데이터 잡음(noisy data)이 될 수 있다.

3) 한 환자당 한 검체만 포함

유전자 변이 조합 추출 방법으로 활용한 Frequent Itemset Mining은 함께 자주 등장하는 아이템의 집합을 특정 최소 빈도를 기준으로 추출하는 방식이다. 유전자 변이 데이터 중 같은 환자의 검체가 존재하면 조합 추출 시 문제가 발생할 수 있다. 같은 환자의 같은 암종에서 추출한 검체를 비교 분석할 시 그 유전자 변이 양상이 매우 흡사하여 데이터 잡음(noisy data)이 될 수 있기 때문이다.

예를 들어 100명이 참여한 임상 데이터에서 최소 빈도를 0.03으로 설정하였을 때, 3명 이상이 가지고 있는 조합이 추출된다. 이때 같은 환자의 검체가 2개 혹은 전체인 3개일 때 추출된 조합은 유효한 조합이라고 볼 수 없다. 또한, 최종적으로 유전자 변이 조합과 치료 반응의 상관성을 계산할 때 해당 조합을 가진 환자의 수가 치료 반응에 따라 더욱 세분되어 적어지므로 1, 2명이 매우 큰 차이가 될 수 있다. 따라서 한 환자당 여러 검체가 있을 시 사전에 파악하여 데이터 필터링을 해야 한다. 환자

/검체 데이터에 '환자 당 샘플 수' 데이터 컬럼이 있는 경우 구분이 용이하다. 일부 데이터의 경우 환자 ID의 뒤에 일정한 규칙의 숫자 혹은 문자를 첨가하여 검체 ID를 생성하기 때문에 이를 활용해 구분할 수 있다. 또는 '1 patient 1 sample', 한 환자당 한 검체만 기록하였다는 조건이 명시되어있는 경우 이 데이터 전처리 작업을 진행하지 않았다.

나. 검체 ID - 유전자 변이 목록 생성

임상 데이터 중 환자/검체 메타데이터, 유전체 변이 주석 형식(MAF) 데이터를 활용해 한 환자의 검체가 가진 변이 유전자 목록을 추출하는 작업이다. 변이가 있는 유전자는 유전자 명(HGNC Symbol)으로 표기하고, 복제수 변이(copy number alteration)가 있는 유전자는 HGNC Symbol 뒤에 AMP(Amplification) 혹은 DEL(Deletion)을 표기하였다.

단순 유전자 명 표기로는 실제 유전자 변이로 인해 발생할 수 있는 단백질 변이를 추측하기 어렵다는 한계가 있으나, 단백질 변이 단위까지 반영하게 되면 환자간 유전자 변이의 일치도가 현저히 떨어져 조합을 추출하기 어렵다. 또한, 아이템 수의 증가로 인해 연산 시간이 급격히 증가하게 되어 조합 추출 시 시간적 효율성이 떨어진다. 따라서 유전자 명을 기준으로 조합을 추출하고, 추후 유의한 조합을 가진 환자들의 데이터를 대조하여 유전자 변이 주석 형식(MAF) 데이터에서 필요한 데이터를 선별 및 확장하여 분석에 활용할 수 있도록 하였다. 유전자 변이 조합을 단백질 단위까지 확장하는데는 'HGVSp_Short'[18] 데이터 컬럼이 사용될 수 있다.

그림 4. 검체-유전자 변이 목록 변환에 사용되는 데이터와 전처리 완료 결과

CLINICAL PATIENT/SAMPLE DATA

patient_id	sample_id	RECIST	oncotree_code	sequence_assay_id	cancer_type	cancer_type_detailed
98	98	PR	IDC	MSK IMPACT 410	Breast Cancer	Breast Invasive Ductal Carcinoma

MUTATION ANNOTATION FORMAT

Subject_ID	Hugo_Symbol	Entrez_Gene_Id	Chromosome	Start_Position	End_Position	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	HGVSp_Short
98	ERBB2	2064	17	37851962	37851963	In_Frame_Ins	INS	-	-	GCCTCCCCA	p.P778_P780dup
98	ARID1A	8289	1	27087503	27087503	Nonsense_Mutation	SNP	C	C	T	p.R89*
98	BMPR1A	657	10	88677065	88677065	Missense_Mutation	SNP	C	C	T	p.R284C
98	NF1	4763	17	29562734	29562734	Nonsense_Mutation	SNP	C	C	T	p.Q1272*
98	NF1	4763	17	29508508	29508509	Splice_Site	INS	-	-	T	p.X218_splice

Copy Number Alteration Data

SAMPLE_ID	HGNC_SYMBOL	cna_value
98	AKT3	2
98	CDC73	2
98	CDKN1B	2
98	FH	2
98	GATA3	2
98	H3F3A	2
98	HIST3H3	2
98	IKBKE	2
98	IL10	2
98	MCL1	2
98	MDM4	2
98	PARP1	2
98	PLCG2	-2
98	TNFRSF14	2
98	ZFHX3	-2

AKT3_AMP
CDC73_AMP
CDKN1B_AMP
FH_AMP
GATA3_AMP
H3F3A_AMP
HIST3H3_AMP
IKBKE_AMP
IL10_AMP
MCL1_AMP
MDM4_AMP
PARP1_AMP
PLCG2_DEL
TNFRSF14_AMP
ZFHX3_DEL

Data Integrate & Simplify

98 Breast ZFHX3_DEL GATA3_AMP HIST3H3_AMP MCL1_AMP ARID1A CDC73_AMP PLCG2_DEL FH_AMP MDM4_AMP IKBKE_AMP TNFRSF14_AMP CDKN1B_AMP ERBB2 PARP1_AMP NF1 H3F3A_AMP IL10_AMP BMPR1A AKT3_AMP

1) 유전체 변이 주석 형식 데이터

유전체 변이 주석 형식(Mutation Annotation Format, MAF)[19]이란, VCF(Variant Call Format) 파일로부터 추출한 유전체 변이 데이터를 통합하여 TSV 형태로 정리한 파일 형식을 말한다. 미국 국립 보건원(NIH, National Institutes of Health) 산하의 국립 암 센터 (NCI, National Cancer Institute)에서 주관하는 유전적 데이터 공유지(GDC, Genomic Data Commons)에서 정의하였으며, 암 유전체 아틀라스(TCGA, The Cancer Genome Atlas Program)[3]에서 사용한 유전체 변이 주석 형식에서 몇 가지를 추가한 형태이다. 표현 가능한 데이터 형식은 126개가 있다. 본 연구에서는 HUGO 유전자 명명 위원회(HUGO Gene Nomenclature Committee)[20]에서 합의한 유전자 명인 HGNC_ID (혹은 HGNC_Symbol)와 검체/환자 ID를 추출하여 사용하였다.

2) 환자/검체 메타 데이터

환자와 검체(sample)에 대한 메타데이터를 말한다. 메타데이터란 데이터를 설명할 수 있는 데이터이다. 임상 데이터마다 차이는 있으나, 주로 환자 ID, 검체 채취 시 환자의 나이, 인종 등이 기록되어있다. 검체 메타데이터는 암종(cancer type), 세부 암종(specific cancer type) 암종 축약 코드(oncotree code)[21], 초발/전이암 분류, 초발/전이암 세부 분류, 검사 방법의 고유 ID, 암 유형, 세부 암 유형 등이 기록되어있다. 환자/검체 메타데이터를 통합 제공하기도 한다. 치료 반응데이터는 주로 검체 메타데이터에 포함되어있다. 본 연구에서는 검체 ID, 암종, 암종 축약 코드를 추출하여 사용하였다.

3) 이산화된 복제수 변이 데이터

복제수 변이(CNA, Copy Number Alteration)란 유전체의 구조적 변이(Structural variation) 중 하나로, 염색체의 특정 구역이 복사 또는 삭제되는 것을 말한다. 임상 데이터마다 그 배포 형태가 상이하다. 유전자 명칭과 Amplification, Deletion을 직접 표기하거나, 유전자 명칭과 숫자로

된 복제수 값을 표기하는 경우 등이 있다. 복제수 값은 주로 NA(값 없음), -2, -1, 0, 1, 2 값으로 표기되며 의미하는 바는 표 3과 같다. 이 중 2 (high-level amplification), - 2(deep loss)를 각각 Amplification, Deletion이라고 한다. 본 연구에서는 다양한 형태의 CNA 데이터를 통일하여 유전자 명(HGNC symbol)에 AMP(amplification), DEL(deletion)을 덧붙여 표기하였다(표 4).

표 3. 이산화된 복제수 변이 데이터의 값과 그 의미

CNA 값	의 미
NA	값 없음
-2	deep loss, possibly a homozygous deletion
-1	single-copy loss
0	diploid
1	low-level gain
2	high-level amplification

표 4. 복제수 변이 데이터 표기 예시

CNA 값	표기 형식	예 시
-2	유전자명_DEL (HGNC_Symbol_DEL)	CDH1_DEL
2	유전자명_AMP (HGNC_Symbol_AMP)	PAK1_AMP

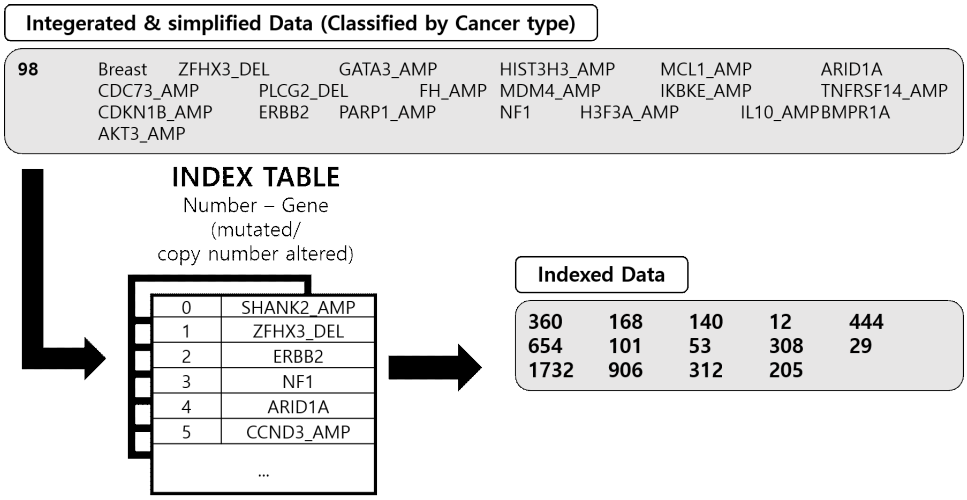
3. 유전자 변이 조합 추출

전처리가 완료된 한 검체 ID-유전자 변이 목록을 이용해 유전자 변이 조합을 추출한다. 항목의 목록을 바탕으로 함께 자주 발생하는 항목 조합을 발생 빈도를 기준으로 추출하는 Frequent Itemset Mining 방법을 사용하였다. 많은 연산량을 빠른 시간 안에 계산할 수 있는 GPU 기반의 Frequent Itemset Mining 도구인 G-Miner[22]를 사용하였다. 환자별 유전자 변이 목록을 해당 도구의 입력 데이터로 사용하기 위한 데이터 인덱싱을 거친 후, 최소 추출 빈도인 지지도(support)를 기준으로 조합을 추출하였다. 이후 인덱스값으로 구성된 유전자 변이 조합과 지지도를 유전자 명인 Hgnc symbol로 맵핑하여 최종적으로 유전자 변이 조합과 지지도의 목록을 추출하였다.

가. 데이터 인덱싱

조합 추출 도구인 G-Miner[22]의 입력 데이터로 사용하기 위해 각 개체를 수 형태의 데이터(numeric data)로 맵핑하는 작업이다. 작업의 시간적 효율을 위해 모든 유전자 변이 개체를 맵핑한 인덱스 테이블을 생성하여 반복적으로 사용하였다. 숫자로 맵핑이 완료된 결과는 그림 5와 같다. 각 검체가 가진 변이 유전자가 인덱싱된 값은 공백을 기준으로, 검체/환자는 개행문자로 구분된다.

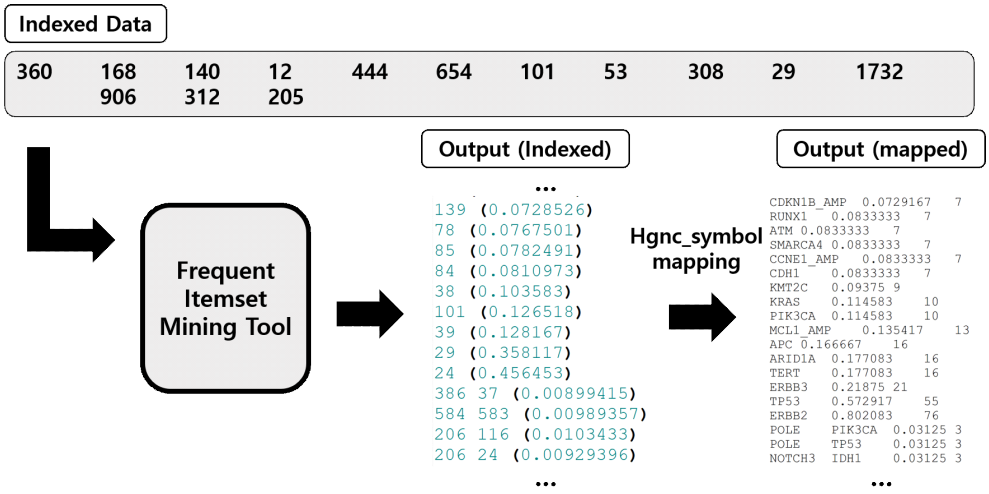
그림 5. 데이터 인덱싱 과정 및 결과



나. 유전자 변이 조합 추출

G-Miner는 대용량 데이터를 대상으로 하여 GPU 기반으로 Frequent Itemset Mining을 시행할 수 있는 도구이다[22]. 입력 데이터로 데이터 인덱싱이 완료된 검체 별 유전자 변이 목록을 사용하였다. 조합 추출 시 최소 발생 빈도인 지지도(support)를 설정해야 하며, 이 값을 기준으로 높은 빈도로 관측되는 조합만이 결과로 출력된다. 지지도는 임상 데이터의 특성과 도구의 실행 환경에 맞추어 다르게 설정할 수 있다. 본 연구에서는 대상 환자가 100명 이하인 경우 적은 데이터에서 다양한 조합이 추출될 수 있도록 3명 이상에게 발견된 조합이라면 추출될 수 있도록 하였다. 대상 환자 수가 50명일 경우 지지도 값은 0.06로 설정하게 된다. 환자 수가 100명 이상이면 지지도 값을 0.03으로 통일하여 진행하였다. 유전자 변이 조합은 인덱싱된 형태로 추출된다. 이를 인덱스 테이블을 거쳐 유전자 명으로 변환하여 최종적으로 변이 유전자 조합과 지지도(support) 값, 해당하는 실제 환자 수를 추출하였다.

그림 6 유전자 변이 조합 추출 과정 및 결과



다. 유전자 변이 조합과 환자/검체 ID 맵핑

추출된 유전자 변이 조합을 환자들의 유전자 변이 데이터와 대조하여, 해당 유전자 변이를 가진 환자/검체들의 ID를 추출한다. 결과 데이터로 유전자 변이 조합과 해당 조합을 가진 환자/검체 ID의 목록의 쌍(pair)이 추출된다. 한 환자/검체는 여러 유전자 변이 조합에 맵핑될 수 있다. 치료 반응과 유의한 상관성을 보이는 유전자 변이 조합을 가진 환자/검체를 심화 분석할 때 해당 ID를 활용할 수 있다.

4. 유전자 변이 조합과 치료 반응 데이터 맵핑

가. 항암 치료 반응 지표

본 연구에서는 항암 치료 반응데이터로 주로 RECIST 1.1 [11]를 활용하였으며, 다양한 항암 치료 반응데이터의 활용 가능성을 보여주기 위해 추가적으로 수술 후 재발 여부를 의미하는 DFS(disease free status)를 활용하였다. RECIST 1.1 외 사용되는 항암 치료 반응 지표는 반드시 이진 형태(binary)여야 한다. RECIST의 경우 4가지 지표를 치료 반응이 좋은 그룹(CR, PR), 치료 반응이 나쁜 그룹(PD, SD)로 분리하여 활용하였다. 본 연구에서 활용한 DFS의 경우 재발함/재발하지 않음으로 분리되어있다.

1) RECIST 1.1

Response Evaluation Criteria In Solid Tumors의 약자로, 고형 종양(solid tumor)의 항암 치료 요법 시 그 반응성을 객관적으로 평가할 수 있는 지표로 널리 사용되고 있다. 1981년 세계보건기구(WHO)에서 사용되던 종양 반응 평가 지침을 기반으로 2000년 처음 발표되었으며, 이를 수정 및 보완한 RECIST 1.1 버전이 2009년 발표되었다. 본 연구에서는 RECIST 1.1을 활용한 데이터만을 사용하였다. 4개의 종양 반응 지표가 있으며 CR(Complete Response), PR(Partial Response), SD(Stable Disease), PD(Progressive Disease) 총 4가지로 나뉜다[11]. 본 연구에서는 CR, PR을 치료 반응이 있는 그룹, SD, PD를 치료 반응이 없는 그룹으로 분류하였다. 면역 항암 치료 반응 평가에 사용될 수 있는 종양 반응 평가 지표인 iRECIST는 RECIST 1.1에 기반하고 있으나, 종양 반응 지표가 5개이며 평가 기준 또한 다소 다르므로 본 연구에서는 RECIST 1.1를 사용한 데이터만을 사용하였다.

표 5. RECIST 1.1 지표

분 류	축약 표기	의 미	기 준
치료 반응 있음	PR	Partial Response	표적 병변의 지름의 합이 30% 이상 감소
	CR	Complete Response	표적 병변이 모두 사라짐
치료 반응 없음	PD	Progressive Disease	표적 병변의 지름의 합이 20% 이상 증가, 혹은 새로운 병변이 한 개 이상 나타남
	SD	Stable Disease	표적 병변의 변화가 크게 없어 PR/PD로 분류될 수 없음

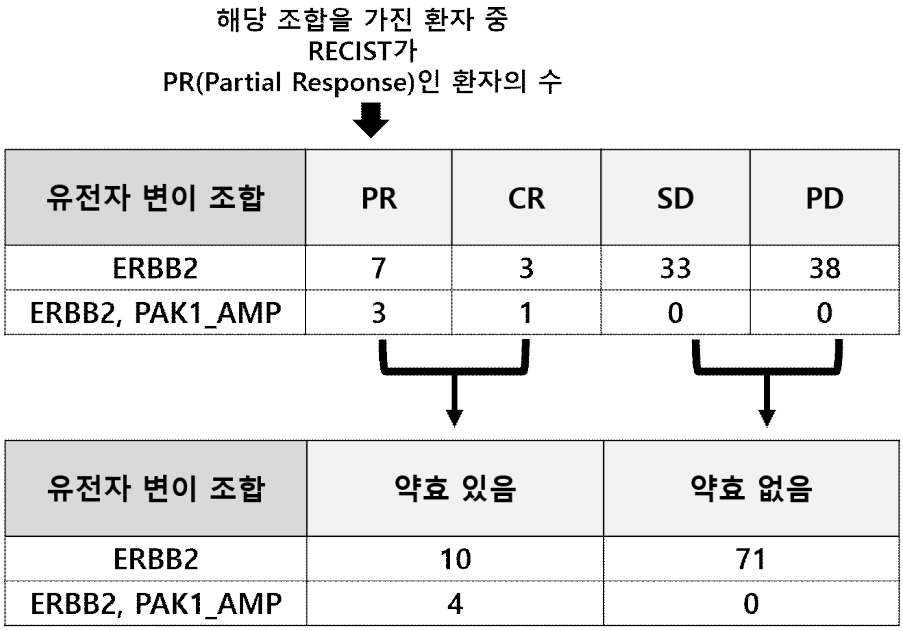
2) 이진 형태 항암 치료 반응 지표

지표가 정량적인 값이 아닌 치료 반응 있음/없음 으로 구분할 수 있는 지표라면 RECIST를 대체하여 활용할 수 있다. 그 활용 예시를 보여주기 위해 간 절제술 이후 간 세포암의 재발 여부를 0 혹은 1로 표현한 DFS (Disease Free Status) 값을 활용하였다. 해당 임상 데이터에서 0은 수술 후 재발하지 않음, 1은 재발하였음을 의미한다.

나. 유전자 변이 조합과 치료 반응데이터 맵핑

추출된 유전자 변이 조합과 두 그룹으로 분류된 치료 반응데이터를 맵핑하는 과정이다. 해당 유전자 변이 조합을 가진 환자들의 치료 반응데이터를 취합하여 각 그룹에 값을 가산하는 방식이다. RECIST의 경우, 특정 유전자 변이 조합을 가진 환자의 치료 반응이 CR 혹은 PR이었다면 치료 반응이 있는 그룹에 가산하고, PD 혹은 SD일 경우 치료 반응이 없는 그룹에 가산한다. 이 과정을 모든 조합과 환자에 대해 반복하며 진행한다. 유전자 변이 조합별로 해당 조합을 가진 환자 중 치료 반응이 있는 환자는 몇 명이었는지, 치료 반응이 없는 환자는 몇 명이었는지 셈하는 과정이다. 이진 형태 항암 치료 반응 지표의 경우 마찬가지로 각 그룹에 값을 가산하였다.

그림 7 유전자 변이 조합을 가진 환자와 치료 반응 맵핑 예시 (RECIST)



유전자 변이 조합	PR	CR	SD	PD
ERBB2	7	3	33	38
ERBB2, PAK1_AMP	3	1	0	0

유전자 변이 조합	약효 있음	약효 없음
ERBB2	10	71
ERBB2, PAK1_AMP	4	0

5. 유전자 변이 조합별 치료 반응과의 유의도 계산

가. 베이즈 요인

베이즈 요인은 대립하는 두 가설 중 데이터가 어떤 가설을 더 지지하는지에 대한 값을 정량적으로 가늠할 수 있도록 돕는 지표이다. 본 연구에서는 귀무가설(H1)을 '이 유전자 변이 조합을 가진 환자에게는 해당 항암 치료 요법의 치료 반응이 없다.'로 설정하였다. 이에 대립하는 대립가설(H0)은 '이 유전자 변이 조합을 가진 환자에게는 해당 항암 치료 요법의 치료 반응이 있다.' 즉, 치료 요법의 효과가 있는 것으로 설정하였다.

데이터가 각 가설을 얼마나 지지하는지에 대한 값을 관측된 값을 기반으로 계산한다. 귀무가설(H1)의 경우, 치료 반응이 없는 전체 환자 수 중 해당 유전자 변이 조합을 가진 환자 수의 비율을 계산한다 (Algorithm 1의 N_{nb}). 대립가설(H0)의 경우, 치료 반응이 있는 전체 환자

수 중 해당 유전자 변이 조합을 가진 환자 수의 비율을 계산하였다 (Algorithm 1의 N_b). 이때 N_{nb}과 N_b 중 어떤 값이 큰지에 따라, 최종 베이즈 요인 값이 유의성을 가질 때 데이터가 어느 가설을 더 지지하는지 방향성이 결정된다. 즉, 해당 유전자 변이 조합이 치료 반응과 유의한 상관성이 있을 때, 해당 조합이 치료 반응이 있는 쪽으로 관계가 있는지, 치료 반응이 없는 쪽으로 관계가 있는지 판가름하는 기준이 된다.

나. 베이즈 요인 계산

본 연구에서 사용한 베이즈 요인 계산 방법[23]을 알고리즘 형식으로 작성하였다. (Algorithm 1) 치료 반응이 있는 전체 환자 수 P_b, 치료 반응이 없는 전체 환자 수 P_{nb}, 특정 유전자 변이 조합 C, 해당 조합을 가지면서 동시에 치료 반응이 있는 환자의 수 CP_b, 해당 조합을 가지면서 동시에 치료 반응이 없는 환자의 수를 CP_{nb}를 입력값으로 취한다. 출력값은 해당 유전자 변이 조합 C와 치료 반응 간의 유의성을 계산한 베이즈 요인 값 BF이다.

해당 유전자 변이 데이터의 귀무가설 지지 여부를 계산하기 위해 정규화된 값 N_b, N_{nb}를 사용하여 임의변수 tendency에 값을 입력하게 된다. N_{nb} 값이 큰 경우 데이터는 귀무가설을 지지하며(support) 이는 곧 베이즈 요인이 유의성을 가질 때, 해당 유전자 변이 조합을 가진 환자들의 데이터는 치료 반응이 없다는 가설을 지지함을 의미한다. N_b 값이 큰 경우 데이터는 귀무가설을 지지하지 않으며(reject) 이는 곧 유의한 베이즈 요인 값일 때 해당 유전자 변이 조합을 가진 환자들의 데이터는 치료 반응이 있다는 가설을 지지함을 의미한다.

해당 유전자 변이 조합을 가진 환자들이 치료 반응이 있는 경우, 없는 경우의 기대 빈도를 계산한다. 그 후 각 기대 빈도가 0인 경우와 해당 유전자 변이 조합을 가진 환자 중 치료 반응이 있는 환자의 수, 없는 환자의 수인 CP_b, CP_{nb}값이 0인 경우가 아니라면, 기대 빈도 CP_b, CP_{nb}을 통해 자연로그 값을 구한다. 자연로그 값을 기반으로 로그 우도(log likelihood)를 계산하고, 최종적으로 베이즈 요인 값을 도출한다.

Algorithm 1. Bayes Factor Calculation

Input:

C : The set of mutated genes $\{G1, G2, G3...\}$

P_b : The number of patients with cb (clinical benefit)

P_{nb} : The number of patients with nb (no clinical benefit)

CP_b : The number of patients with cb whose harboring C

CP_{nb} : The number of patients with nb whose harboring C

output:

BF : Bayes Factor

1. **procedure**
2. *// calculate tendency whether the data support or reject null hypothesis (no clinical benefit) using normalized value(N)*
3. $N_b = \frac{CP_b}{P_b}$
4. $N_{nb} = \frac{CP_{nb}}{P_{nb}}$
5. **if** $N_{nb} > N_b$
6. **then** tendency ← "support"
7. **else** tendency ← "reject"
8. *// calculate expected frequencies (ef)*
9. $ef_b \leftarrow P_b \times \frac{CP_b + CP_{nb}}{P_b + P_{nb}}$
10. $ef_{nb} \leftarrow P_{nb} \times \frac{CP_b + CP_{nb}}{P_b + P_{nb}}$
11. *// calculate log likelihood (logL)*
12. **if** $ef_b = 0$ or $CP_b = 0$
13. **then** $\ln_b \leftarrow 0$
14. **else** $\ln_b \leftarrow \ln(\frac{CP_b}{ef_b})$

15. **end if**
16. **if** $ef_{nb} = 0$ or $CP_{nb} = 0$
17. **then** $\ln_{nb} \leftarrow 0$
18. **else** $\ln_{nb} \leftarrow \ln(\frac{CP_{nb}}{ef_{nb}})$
19. $\log L = 2 \times ((CP_b \times \ln_b) + (CP_{nb} \times \ln_{nb}))$
20. **end if**

21. *// calculate bayes factor*
22. $BF = \log L - \ln(P_b + P_{nb})$
23. **end procedure**

베이즈 요인 값의 범위에 따라 특정 유전자 변이 조합과 치료 반응 간의 상관성이 얼마나 유의한지를 나눌 수 있다. 베이즈 요인 값이 음수일 때 유의성이 없는 것으로 간주한다. 베이즈 요인 값이 0 이상 2 이하일 때는 비교적 유의성이 약하다. 2 이상 6 이하일 경우 유의성이 있다고 해석할 수 있지만, 그 유의성이 강하지 않다. 6 이상 10 이하일 경우 강한 유의성을 보인다고 볼 수 있으며, 10 이상일 경우 매우 강한 유의성을 보임을 의미한다. 본 연구 결과에서 유의한 조합 값은 베이즈 요인이 2 이상인 조합을 의미한다.

표 6. 베이즈 요인 값의 범위에 따른 유의성 판단 기준

베이즈 요인 값	유 의 성
– (NEGATIVE value)	non significant
0-2	weak
2-6	moderate
6-10	strong
>10	very strong

모든 조합과 치료 반응에 대한 베이즈 요인 계산이 끝나면, 베이즈 요인 값이 음수인 조합은 유의하지 않으므로 모두 제거한다. 이후 유전자 변이 조합간의 관계가 진부분집합이면서, 해당 환자 목록이 완전히 같은 포함 관계의 조합을 찾아내 제거한다. 예를 들어 유전자 변이 조합 C1, C2 이고, C1 = {GENE A, GENE B} 이고 C2 = {GENE A, GENE B, GENE C} 일 때, C1을 가진 환자의 목록 P1 = {Paitent 1, Patient 2}, C2를 가진 환자의 목록 P2 = {Paitent 1, Patient 2} 이라고 가정하자. 이는 유전자 A, B 변이를 가진 환자가 모두 유전자 C 변이를 가진 것이므로 조합 C1은 C2의 진부분집합이 된다. 따라서 C1은 조합 목록에서 데이터에서 제거된다. 진부분집합 제거 과정을 거친 후 최종적으로 베이즈 요인을 기준으로 내림차순으로 정렬하여 랭킹된 목록을 추출하여 결과로 제공한다.

베이즈 요인 값이 유의할 때, 데이터의 가설 지지 방향성에 따라 해석을 달리할 수 있다. 가설 지지 방향성은 Algorithm 1.에서 변수 tendency이다. 가설 지지 방향성이 'support'이면 귀무가설을 지지함을 의미한다. 이때 베이즈 요인 값이 유의한 범위 안에 들면, 이는 해당 유전자 변이 조합을 가진 환자들이 치료 반응이 없었다고 해석할 수 있다. 반대로 가설 지지 방향성이 데이터가 'reject'이면 귀무가설이 아닌 대립가설을 지지하는 것으로, 이는 해당 유전자 변이 조합을 가진 환자들이 치료 반응이 없지 않았다, 즉 치료 반응이 있다고 해석할 수 있다. 실제 결과 파일에는 치료 반응이 있는 조합은 + (positive), 치료 반응이 없는 조합은 - (negative)로 표현하였다.

결 과

본 장에서는 해당 연구의 결과물 및 활용 예시를 설명한다. 본 연구에서는 다양한 항암 치료 요법 임상 데이터 적용 사례 검증을 위해 3가지 항암 요법, 5개의 데이터를 활용하였다. 표적 치료 요법으로는 2가지 종류의 데이터를 사용하였으며, 표적 단백질과 표적 치료제는 각각 ERBB2-neratinib [12], PIK3CA-alpelisib [13] 이었다. 면역 체크포인트 치료의 경우 첫 번째 데이터[15]는 CTL4, PD-L1, PD-1 억제제, 두 번째 데이터[16]는 PD-L1, PD-1 억제제를 사용한 결과이다. 또한, 수술 이후 암 재발률과 유전자 변이 데이터 간의 상관성을 관측한 간 절제술 임상 데이터[17]를 사용하였다.

풍부한 유전자 변이 조합 추출을 위해 전체 환자 수가 50명 이상인 데이터를 선택하였으며, 데이터 전처리 과정 중 치료 반응데이터가 없거나 유전자 변이 데이터가 없는 환자를 제외하였다. 표 7의 '대상 환자 수'는 전처리 완료 후의 환자 수를 의미한다. 본 연구방법의 '2-가. 데이터 필터링' 기준에 따라 표적 치료① 데이터에서는 141명 중 36명, 표적 치료② 데이터에서 51명 중 8명, 면역 체크포인트 치료①에서 245명 중 0명, 면역 체크포인트 치료② 245명 중 5명, 간 절제술 치료 236명 중 5명이 분석 과정에 포함되지 않았다.

표 7. 활용된 임상 데이터와 대상 환자 수

치료 요법 종류	치료 요법 상세	전체 환자 수	분석 대상 환자 수
표적 치료①	ERBB2 kinase inhibitor (neratinib)	141	105
표적 치료②	PIK3CA inhibitor (alpelisib+AI)	51	43
면역 체크포인트 치료①	CTLA4-inhibitor, PD-L1 inhibitor, PD-1 inhibitor	245	245
면역 체크포인트 치료②	PD-L1 inhibitor, PD-1 inhibitor	245	240
간 절제술	Hepatic resection	236	231

본 도구 활용 시 추출되는 데이터는 해당 항암 치료 반응과 유의미한 상관성을 보이는 랭킹된 유전자 변이 조합의 목록이다. 또한, 각 유전자 변이 조합별 베이즈 요인 값과 그 값이 의미하는 유의성의 정도, 해당 유전자 변이 조합을 가진 환자/검체의 ID와 암종이 함께 제공된다. 사용자는 환자/검체 ID를 통해 원본 MAF 데이터[19] 및 환자/검체 메타데이터에 접근할 수 있다. 예를 들어, 해당 유전자 변이 조합의 단백질 변이 양상(HGVSp_short)[18], 조합을 가진 환자의 암종 분포(cancer type) 등의 심화 분석을 할 수 있다.

1. 표적 치료법 임상 데이터① 분석 결과

본 연구에서 데이터로 사용한 표적 치료 임상 데이터는 ERBB2, ERBB3 유전자 변이가 있는 환자를 대상으로 ERBB2 family 억제제인 neratinib을 투여한 것이다[12]. 유의한 상관성을 보이는 유전자 변이 조합은 2개가 추출되었다. 모두 치료 반응이 있는 그룹으로, 치료 반응이 없는 조합은 추출되지 않았다(표 8-1). 본 임상 데이터는 모든 환자가 ERBB2 혹은 ERBB3 유전자 변이가 있으므로 추출된 조합은 모두 ERBB2/ERBB3를 포함하고 있다. PAK1 유전자 증폭 변이, ERBB2 변이 조합이 베이즈 요인 값이 13.53 값으로 유의성이 가장 높은 조합으로 추출되었다. 두 번째로는 11번 유전자에 잇달아 배치되어있는 11q13.3 증폭 변이 (FGF3, FGF4, FGF19, CCND1)와 ERBB2 변이가 함께 등장하는 조합이었으며 베이즈 요인 값은 2.72으로 1위 조합보다 약한 유의성을 보였다.

표 8-1. 표적 치료 임상 데이터① 에서 추출된 조합 (POSITIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	PAK1_AMP, ERBB2	13.53	very strong
2	FGF3_AMP, FGF4_AMP, FGF19_AMP, CCND1_AMP, ERBB2	2.72	moderate

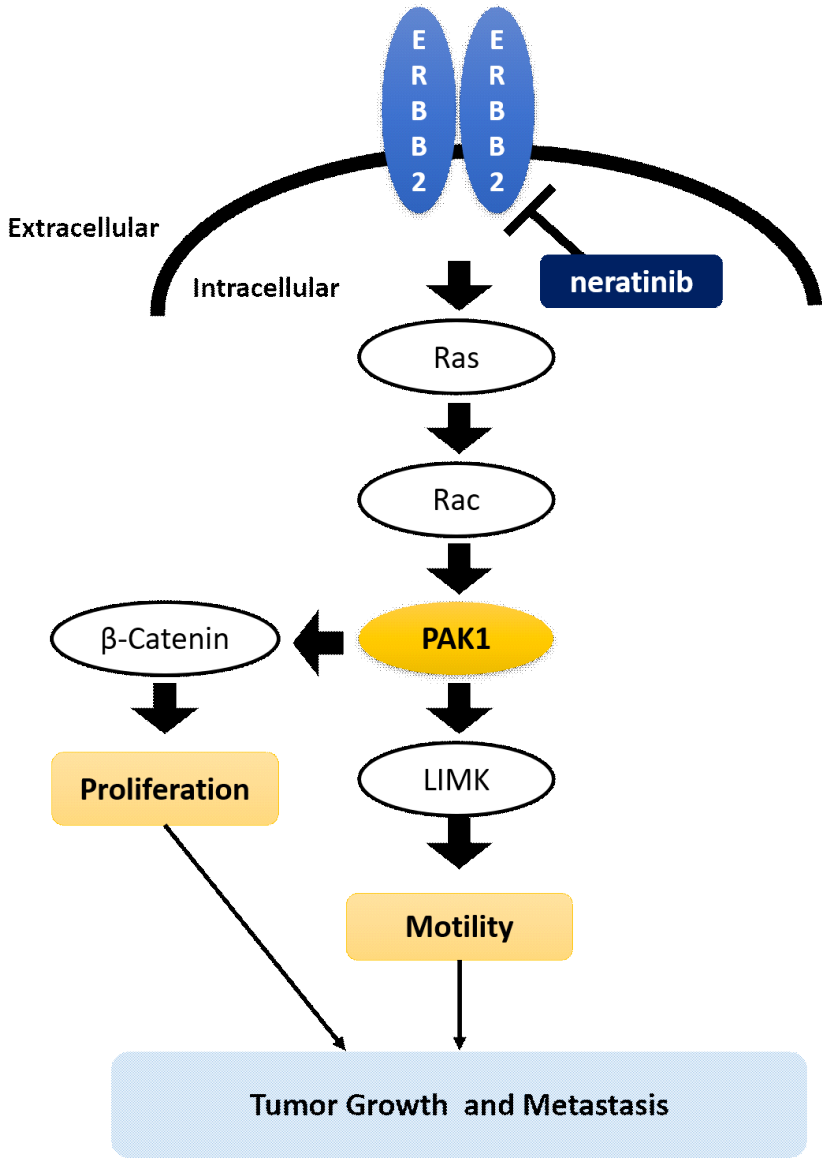
1) PAK1_AMP, ERBB2 조합과 neratinib의 상관성 (POSITIVE)

본 결과는 ERBB2 변이와 PAK1 증폭 변이 조합을 가진 환자에게 ERBB2 억제제인 neratinib을 투약하였을 때 치료 반응이 있었다는 것을 의미한다. Arias-Romero, L. E. et al. (2011)[24]에 따르면, ERBB2 변이로 인해 정상 유방 상피세포가 암세포로 변형되는 과정에 있어 PAK1이 매우 중요한 역할을 한다. 저자들은 xenograft 실험을 통해 ERBB2와 PAK1가 암세포의 증식과 전이에 있어 함께 주요한 역할을 하는 것을

증명하였다. PAK1 억제제를 투여하였을 때 ERBB2로 인한 유방암 세포의 암 변형이 저지되고, 이에 더하여 PAK1 억제제와 ERBB2 억제제를 동시에 투여하였을 때 암을 저지하는 효과가 더욱 큰 것을 확인하였다.

이는 ERBB2와 Rac/Pak 패스웨이의 관계로 설명할 수 있다[25]. ERBB2 변이와 PAK1 증폭 변이가 있는 암세포의 경우, ERBB2의 변이로 인한 downstream 여파, PAK1의 증폭으로 인한 Rac/Pak 패스웨이의 과도한 활성화가 해당 암세포의 메커니즘에 주요한 역할을 하고 있음을 가정할 수 있다. 이때 ERBB2 억제제를 투여하게 되면 ERBB2로 인한 downstream의 신호 전달을 막게 된다. 이로 인해 Rac/Pak 패스웨이로 인한 암세포의 활동이 저지될 것이므로, 해당 변이 양상을 가진 암세포가 ERBB2 억제제에 취약한 이유를 설명할 수 있다.

그림 8. ERBB2, PAK1 Pathway와 neratinib의 작용



2. 표적 치료법 임상 데이터② 분석 결과

해당 데이터는 호르몬 수용체 양성(HR+) 유방암 환자를 대상으로 PIK3CA inhibitor인 alpelisib과 항 에스트로겐(antiestrogen) 치료제인 아로마타아제 억제제 (Aromatase inhibitor)를 함께 투여한 결과이다[13]. 총 51명의 환자를 대상으로 하였으며 유전자 시퀀싱 분석방법으로 ctDNA, 종양 분석을 병용하였다. 투약 전과 투약 후의 데이터를 따로 제공하나, 본 연구에서는 투약 전 데이터만을 활용하였다. RECIST가 없는 8명의 환자를 제외하였으며, ctDNA와 종양 분석 데이터를 모두 가진 환자의 경우 두 가지 분석 결과를 통합하여 입력 데이터로 활용하였다. 베이즈 요인 값이 2 이상인 유전자 변이 조합은 없었으나, 베이즈 요인 값 1.2의 조합이 1개 추출되었다. 해당 조합은 NOTCH1, ARID1A, PIK3CA 변이가 동시에 있는 환자가 alpelisib에 대한 약 반응성이 있었음을 의미한다.

표 8-2. 표적 치료 임상 데이터② 에서 추출된 조합 (POSITIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	NOTCH1, ARID1A, PIK3CA	1.20	weak

1) NOTCH1, ARID1A, PIK3CA 조합과 면역 체크포인트 억제제의 상관성 (POSITIVE)

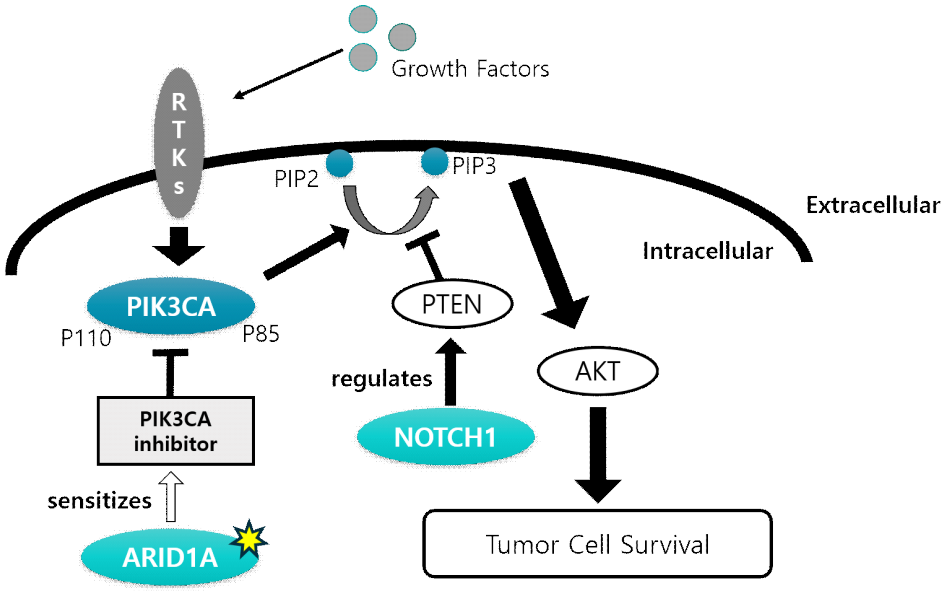
ARID1A와 PIK3CA의 유전자 변이가 동시에 있을 때 세포의 집단 침습(collective invasion)을 촉진한다고 알려져 있다[26]. 집단 침습은 단일 세포의 침습과는 다른 기전을 통해 여러 세포가 한꺼번에 이주하는 것으로, 암의 전이에 있어 주요한 역할을 한다[27]. 또한 ARID1A 기능 소실 변이(loss of function mutation)로 인해 ARID1A의 발현량이 적어진 경우, PIK3CA 억제제의 민감성이 증가한다는 사실이

증명되었다[28]. 이는 ARID1A 기능 소실 변이가 있는 암세포는 그렇지 않은 암세포에 비해 PIK3CA 억제제가 더 잘 들었음을 의미한다.

NOTCH1 유전자는 PI3K/AKT 패스웨이[29]에 간접적으로 영향을 끼친다. PIK3CA는 세포 외부로부터 생장 인자(Growth factor)와 티로신 인산화 효소 수용체(RTKs, Receptor tyrosine kinases)로 인해 신호를 받아 활성화된다. 활성화된 PIK3CA는 PIP2 (phosphatidylinositol(4,5)-bisphosphate)가 PIP3 (phosphatidylinositol (3,4,5)-bisphosphate)로 전환되는 과정에 관여한다. PIP3는 AKT 활성화에 영향을 미쳐 결과적으로 AKT를 통한 암세포의 생존에 영향을 끼치게 된다. 이 때 PTEN 유전자는 PIP2가 PIP3로 전환될 때 억제작용을 하는데, NOTCH1이 PTEN을 조절하는 유전자이다[30].

위와 같은 정보를 종합해보았을 때, NOTCH1, ARID1A, PIK3CA 유전자 변이 조합을 가진 암세포가 PIK3CA 억제제의 좋은 치료 반응의 상관성을 보인 결과에 대해 다음과 같은 분석을 할 수 있다. 암세포에서 NOTCH1 유전자 변이는 주로 기능 증가(gain of function)현상을 보인다는 점을 반영하여 분석하였다[31]. NOTCH1의 기능 증가로 PTEN으로 인한 PIP2-PIP3 전환이 억제되어 PI3K/AKT 패스웨이로 인한 암세포의 기능이 저하된 상태를 가정한다. 이때 PIK3CA 억제제가 투약되면 ARID1A의 변이로 인한 PIK3CA 억제제 민감성이 증가한 상태이므로, PIK3CA 억제제가 암세포에 치명적인 영향을 끼쳤을 것이라고 유추할 수 있다.

그림 9. ARID1A, NOTCH1, PIK3CA 변이가 PIK3CA inhibitor의 치료 반응 민감도에 미치는 작용



3. 면역 체크포인트 치료법 임상 데이터① 분석 결과

본 데이터는 비 소세포 폐암(NSCLC, Non-Small Cell Lung Cancer) 환자를 대상으로 면역 체크포인트 억제제인 PD-1 억제제, PD-L1 억제제를 사용한 결과이다[15]. 240명의 환자에게서 유의한 상관성을 보이는 유전자 변이는 단일 1개, 조합 3개로 총 4개가 추출되었다. 이 중 치료 반응이 있는 조합은 4개, 치료 반응이 없는 조합은 0개였다. 단일 유전자 변이는 POLE이었으며 베이즈 요인 값은 4.17 이었다. POLE 유전자 변이를 가진 11명 중 10명이 TP53 변이가 함께 있었으며, 이는 유의한 조합 중 1위였다.

표 9. 면역 치료 임상 데이터① 에서 추출된 조합 (POSITIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	POLE, TP53	5.92	moderate
2	KRAS, TP53	3.37	moderate
3	NTRK3, TP53	2.75	moderate

1) POLE, TP53 조합과 면역 체크포인트 억제제의 상관성 (POSITIVE)

POLE은 DNA 복제와 복구에 참여하는 효소인 DNA polymerase epsilon의 촉매 소단위(catalytic subunit)를 인코딩하는 유전자이다[32]. 또한 POLE 유전자 변이는 암세포의 높은 유전자 변이율(mutation rates), 높은 TMB와 연관이 있다고 밝혀져 있으며, 면역 체크포인트 억제제의 반응 예측 생물지표(predictive biomarker)로써 주목받고 있다[33].

POLE과 TP53은 DNA 손상 복구 기전에 중요한 역할을 하는 유전자(DNA Damage Response genes)라는 공통점이 있다. 또한, 두 유전자는 모두 세포 주기 중 G1(Gap1)-S(Synthesis) 과정에 관여하는 유전자이다[34]. 이와 같은 정보를 종합해보았을 때 두 유전자 변이를 동시에 가진 암세포의 경우 DNA 복제 시 손상 복구 기전에 문제가 발생하여 유전자 변이 발생률이 높아질 것으로 예측할 수 있다. 높은 유전자 변이율은 TMB과 관계가 있으므로 이를 통해 면역 체크포인트 억제제와의 상관성을 유추할 수 있다. 하지만 두 유전자 변이를 동시에 가졌을 때를 특정한 연구 결과는 존재하지 않으므로, 추가적인 실험을 통한 뒷받침이 필요하다.

2) KRAS, TP53 조합과 면역 체크포인트 억제제의 상관성 (POSITIVE)

Dong, Z. Y. et al. (2017)에 따르면 폐 선암(Lung Adenocarcinoma)종의 면역 체크포인트 치료(PD-1 억제제, PD-L1 억제제)에 있어 KRAS, TP53의 co-occurring mutation이 잠재적인 반응 예측 생물지표가 될 수 있다[35]. RNA-seq 분석, mRNA 발현량 분석, RPPA(Reverse Phase Protein Arrays) 단백질 발현량 분석 결과 TP53, KRAS 유전자 변이가 함께 있는 경우 PD-L1 발현량이 현저히 높음을 발견하였다. 또한, KRAS, TP53 유전자 변이를 함께 가진 경우 TMB 값이 다른 유전자 변이보다 현저히 높았음을 보여주었다. 결과적으로 TP53, KRAS 변이가 PD-L1의 발현, 암의 면역원성(tumor immunogenicity) 증대, T세포 침윤과 유의한 연관이 있으며, PD-1 억제제, PD-L1 억제제를 통한 면역치료 시 해당 유전자 변이 조합을 예측 인자로 활용할 수 있을 것이라 기대하였다[35]. 이는 본 연구 결과에서 추출된 KRAS, TP53 유전자 변이 조합과 면역 체크포인트 억제제의 상관성을 뒷받침하는 결과라고 볼 수 있다.

4. 면역 체크포인트 치료법 임상 데이터② 분석 결과

본 데이터는 면역 체크포인트 억제제로 CTLA4 억제제, PD-1 억제제, PD-L1 억제제를 사용한 결과이다[16]. 이 중 유의한 상관성을 보이는 조합은 총 1656개였다. 이 중 치료 반응이 있는 조합은 1651개, 치료 반응이 없는 조합은 5개였다. 표 10-1은 상관성 값을 기준으로 정렬한 유전자 변이 조합 중 치료 반응이 있는 조합 1651개 중 상위 10개를 나열한 것이다. 이 중 16.60의 베이즈 요인 값을 보인 PAPPA2와 RYR1이 가장 상관성이 높았으며, 상위 10개의 조합 중 7개의 조합이 해당 유전자 변이 조합 외에 1가지가 더 추가된 형태의 조합임을 확인할 수 있었다. 표 10-2의 경우 치료 반응이 없는 조합 5개를 나열한 것으로, NLRP4, SPTA1, LRP1B, PCLO 변이 조합이 가장 높은 값으로 해당 유전자 변이 조합을 가진 환자들이 그렇지 않은 환자에 비해 면역 체크포인트 억제제의 치료 반응이 좋지 않았음을 의미한다.

표 10-1. 면역 치료 임상 데이터② 에서 추출된 조합 TOP 10 (POSITIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	PAPPA2, RYR1	16.60	very strong
2	PAPPA2, HYDIN, RYR1	15.22	very strong
3	PAPPA2, C6, RYR1	15.22	very strong
4	PAPPA2, RYR1, TTN	14.74	very strong
5	PAPPA2, SPHKAP, RYR1	13.20	very strong
6	PAPPA2, RYR1, MUC16	13.00	very strong
7	FAM135B, PAPPA2, RYR1	12.95	very strong
8	PCDH15, RYR1, RP1	12.84	very strong
9	PAPPA2, RYR1, CSMD3	12.83	very strong
10	PCDH15, RYR1	11.47	very strong

표 10-2. 면역 치료 임상 데이터② 에서 추출된 조합 (NEGATIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	NLRP4, SPTA1, LRP1B, PCLO	5.94	moderate
2	SLC22A9, MUC16, TTN	2.42	moderate
3	NBEA, DOCK3, MUC16, TTN	2.42	moderate
4	DNAH1, SPTA1, MUC16, TTN	2.42	moderate
5	KIAA1324L	2.07	moderate

1) PAPPA2, RYR1 조합과 면역 체크포인트 억제제의 상관성(POSITIVE)

면역 체크 포인트 억제제는 높은 TMB(Tumor Mutation burden, 종양 돌연변이 부하)값과 유의미한 상관성을 가진다[5]. TMB 값은 세포가 가진 유전체 돌연변이의 총량을 정량적으로 계산한 값으로, 암세포가 많은 돌연변이를 가지고 있을수록 면역 체크포인트 억제제의 치료 반응이 좋다는 것을 의미한다. PAPPA2, RYR1 유전자 변이는 높은 TMB 값과 유의미한 상관관계가 있다[36]. 이와 같은 정보를 종합해보면, PAPPA2와 RYR1 유전자 변이를 가진 암세포는 높은 TMB 값을 가질 확률이 높고, 높은 TMB 값을 가진 암세포는 면역 체크포인트 억제제의 치료 반응이 더욱 좋을 것으로 예상해볼 수 있다.

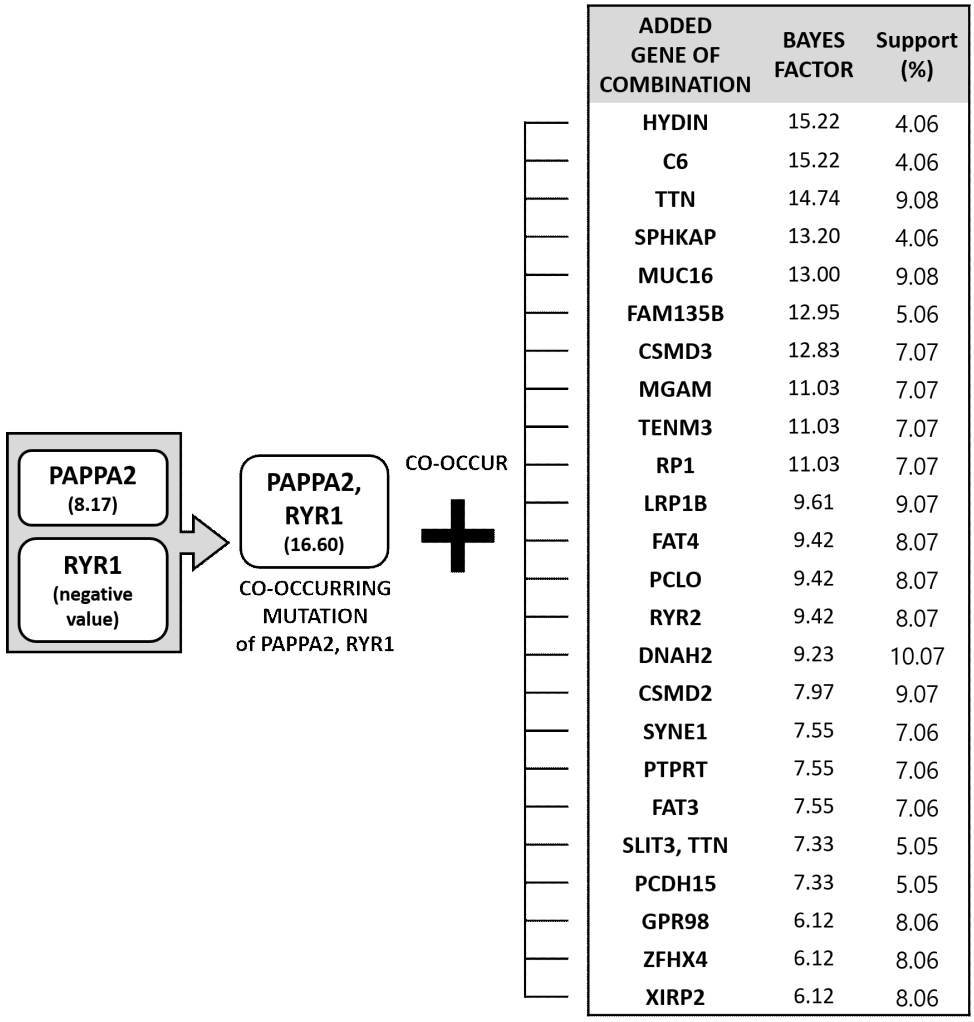
실제로 비 소세포 폐암(NSCLC, Non-Small Cell Lung Cancer) 환자를 대상으로 면역 체크포인트 치료를 시행한 임상 데이터에서 지속적인 치료 반응이 있는 환자 중 4명 이상의 환자가 PAPPA2, RYR1 외 5개 유전자 변이가 있을 확률이 높았으며, 반대로 지속적인 치료 반응이 없는 환자에는 해당 유전자들에 변이가 없다는 것을 발견하였다[36]. 이는 PAPPA2와 RYR1 유전자 변이가 있는 암세포가 면역 체크포인트 억제제에 치료 반응이 있었다는 결과를 뒷받침할 수 있다.

단, 문헌 검색 결과는 PAPPA2와 RYR1의 유전자 변이를 각각 언급한

것으로, 본 연구 결과와는 다소 차이가 있다. 본 연구 결과에서 PAPPA2 유전자 변이의 상관성 값은 8.17으로 강한 유의성을 보였고, RYR1은 음수 값으로 상관성이 없음을 보였다. 두 유전자 변이를 함께 가지는 PAPPA2, RYR1 조합의 상관성 값은 16.60으로 각 단일 유전자 변이보다 훨씬 강한 유의성을 보이는 것을 관측할 수 있다. 이것은 각 단일 유전자 변이가 아닌 두 변이가 함께 일어났을 때 면역 체크포인트 치료 반응이 좋았음을 의미한다. 현시점을 기준으로 두 유전자가 포함된 생물학적 패스웨이 간의 관계와 단백질 간 상호작용(PPI) 연구가 아직 존재하지 않았기 때문에 종합적인 분석은 할 수 없었다. 하지만 이와 같은 결과가 새로운 패스웨이 연구 혹은 단백질 간 관계 연구의 길잡이가 될 수 있을 것으로 기대한다.

그림 10은 상관성 값이 높음 (Bayes Factor > 6)인 조합 중 PAPPA2, RYR1을 포함한 조합 25개를 정렬하고, 해당 조합을 가진 환자 수(support)를 표기한 것이다. HYDIN와 C6 유전자가 PAPPA2, RYR1와 함께 변이가 있을 시 면역 체크 포인트 치료 반응과의 상관성이 가장 높은 것을 확인할 수 있었다. 또한 전체 환자 중 해당 조합을 가진 환자 수를 의미하는 support 값이 높은 유전자 변이 DNAH2, MCU16, TTN (9.08%) 이었다.

그림 10. PAPPA2, RYR1 유전자 변이와 함께 나타나는 유전자 변이 목록



ADDED GENE OF COMBINATION	BAYES FACTOR	Support (%)
HYDIN	15.22	4.06
C6	15.22	4.06
TTN	14.74	9.08
SPHKAP	13.20	4.06
MUC16	13.00	9.08
FAM135B	12.95	5.06
CSMD3	12.83	7.07
MGAM	11.03	7.07
TENM3	11.03	7.07
RP1	11.03	7.07
LRP1B	9.61	9.07
FAT4	9.42	8.07
PCLO	9.42	8.07
RYR2	9.42	8.07
DNAH2	9.23	10.07
CSMD2	7.97	9.07
SYNE1	7.55	7.06
PTPRT	7.55	7.06
FAT3	7.55	7.06
SLIT3, TTN	7.33	5.05
PCDH15	7.33	5.05
GPR98	6.12	8.06
ZFHX4	6.12	8.06
XIRP2	6.12	8.06

2) NLRP4, SPTA1, LPR1B, PCLO 조합과 면역 체크포인트 억제제의 상관성 (NEGATIVE)

치료 반응이 없는 조합의 경우 상관성이 있는 조합은 총 5개가 추출되었다. 그 중 NLRP4, SPTA1, LPR1B, PCLO 유전자 변이 조합을 가진 경우가 다른 조합을 가진 환자들에 비해서 치료 반응이 없었음을 의미한다. 해당 유전자들은 암과 항암요법에 관련된 연구 결과의 수가 적었기

때문에 문헌 검색을 통한 상관성 검증이 어려운 한계가 있었다. NLRP4는 암세포의 자가소화작용(autophagy)의 음성적 조절(negatively regulation)과 관련이 있으며[37], SPTA1는 대장직장암(colorectal cancer), 소세포 폐암(SCLC, Small Cell Lung Cancer)의 발달과 관련이 있는 것으로 알려져 있다[38]. PCLO는 식도 편평 세포암(ESCC, Esphageal Squamous Cell Carcinomas)에서 그 변이가 자주 발견되는 유전자이다[39]. LRP1B는 해당 유전자 변이를 가진 흑색종 피부암(Melanoma)이 높은 TMB와 연관이 있다[40].

비록 현시점의 연구 결과를 바탕으로 NLRP4, SPTA1, LPR1B, PCLO 유전자 변이 조합과 체크 포인트 억제제의 상관성에 대한 문헌을 찾기 어려웠지만, 본 연구 결과를 바탕으로 새로운 생물지표를 위한 실험적 검증을 시도해볼 수 있다.

5. 간 절제술 후 재발 여부 데이터 분석 결과

본 데이터는 간 절제술을 받은 초기 간세포 암 환자 231명을 대상으로 재발 여부를 확인한 결과이다. 추출된 유전자 변이 양상은 총 5개로, 조합 1개, 단일 4개였다. 이 중 4개는 치료 반응이 좋은 그룹, 1개는 치료 반응이 나쁜 그룹이었다. 치료 반응 평가 지표로 RECIST가 아닌 재발 여부를 나타내는 DFS(Disease free status)를 활용하였으며, 값은 재발하지 않았음을 뜻하는 '0:DiseaseFree'와 재발하였음을 뜻하는 '1:Recurred'로 나뉜다.

표 11-1. 간 절제술 데이터에서 추출된 조합 (POSITIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	MUC16, FSIP2	3.40	moderate

표 11-2. 간 절제술 데이터에서 추출된 유전자 변이 (POSITIVE)

랭킹	유전자 변이	베이즈 요인 값	유의성
1	ADAMTSL1	3.40	moderate
2	FANCM	3.40	moderate
3	DYNC2H1	2.47	moderate

표 11-3. 간 절제술 데이터에서 추출된 조합 (NEGATIVE)

랭킹	유전자 변이 조합	베이즈 요인 값	유의성
1	RB1	3.10	moderate

1) RB1 변이와 간 절제술 후 재발 여부의 상관성 (NEGATIVE)

RB1 유전자는 대표적인 종양 억제 유전자(tumor suppressor gene) 중 하나로, 세포 주기의 조절 및 진행에 관여한다. Ahn, S. et al.[17]은 RB1 유전자 변이가 간 절제술 후 재발을 예측할 수 있는 독립적인 생물지표임을 언급하였다. 이들은 해당 데이터의 RB1 변이를 가진 환자들의 RB1 단백질 발현량이 현저히 낮았으며, E2F1(E2 transcription factor 1) 단백질의 발현량은 상대적으로 높음을 발견했다. RB1은 E2F1를 억제하며, E2F1은 RB1으로 인한 자가소화작용과 길항적인 역할을 하는 것으로 알려져 있다[41]. 이는 RB1 변이를 가진 암세포는 자가소화작용이 저해되어있는 상태임을 의미한다. 자가소화작용은 암의 발생(tumorigenesis)을 막는 중요한 기전이다. 위와 같은 기존 연구 결과를 종합해보았을 때, RB1 유전자 변이가 있는 암세포는 세포 자가소화작용이 저해된 상태로 암의 발생을 저지하지 못하며, 이로 인해 간 절제술 이후 암의 재발이 있었음을 유추할 수 있다.

2) MUC16, FSIP2 조합과 간 절제술 후 재발 여부의 상관성 (POSITIVE)

MUC16 유전자는 정상 상피조직의 꼭대기 면(apical surface)에서 보호 역할을 하는 대세포 표면 분자(large cell surface molecule)이다. 정상세포에선 외부 감염을 막는 점막층을 형성하는 데 관여하지만, 몇몇 악성 종양에서도 과발현됨이 발견되었다. 특히 난소암의 초기 발견에 있어 주요한 생물지표로 알려져 있으나, 일반적인 암의 진행과 전이에도 관여한다고 알려져 있다[42]. FSIP2 유전자는 고환 생식 세포 종양(testicular germ cell tumours, TGCT) 환자에게서 증폭 변이가 발견되었지만[43] 다른 암종과의 관계성은 알려지지 않았다. 본 연구 결과에서는 두 유전자 변이가 있는 암세포를 가진 환자는 간 절제술 후 재발이 없었음을 의미한다. 이러한 두 유전자 변이의 조합과 수술 후 재발 여부와의 관계성을 증명하기 위해서는 추가적인 실험적 검증이 필요하다.

고 찰

1. 결과에 대한 고찰

본 연구는 항암 치료 반응에 대한 환자 간 차이를 단일 유전자의 변이로 설명할 수 없는 현상을 유전자 변이 조합의 패턴을 통해 해결하고자 하고자 하였다. 다양한 임상 데이터를 활용해 치료 반응과 상관성이 있는 유전자 변이를 단일을 넘어 2개 이상의 조합을 발굴하는 도구를 개발하였다. 유전자 간의 상호작용 및 패스웨이와 같은 생물학적 맥락 정보를 함께 입력하지 않았음에도, 치료 반응과 관계있는 설명 가능한 유전자 변이 양상을 추출할 수 있었다.

발굴된 유전자 변이 조합의 생물학적 유의성 검증을 직접 문헌 검색 혹은 생물학적 패스웨이 분석을 해야 하는 한계가 있으나, 이는 추가 분석 모듈을 추가함으로써 극복될 수 있다. 유전자 변이 조합을 해당 유전자들이 함께 언급된 문헌 검색, 통합 패스웨이 데이터베이스 탐색을 자동화하여 그 결과를 함께 제공할 수 있을 것이다. 또는 단백질의 생물학적 기능, 분자생물학적 기능, 세포 내 발현 위치 데이터를 제공하는 Gene Ontology와 같은 데이터베이스를 활용하여 유전자 변이 조합 결과를 확장할 수 있다. 이와 같은 추가 분석 모듈을 통해 추후 설명 가능한 모델을 구축할 수 있다.

본 도구를 활용하여 다수의 임상 데이터로부터 상관성 있는 항암 치료 반응과 유전자 변이 양상을 추출하고 이를 통합한다면 항암 요법 치료 반응 데이터베이스를 구축할 수 있을 것이다. 이를 통해 환자의 암세포의 유전자 변이 양상을 기반으로 단일 유전자 변이뿐 아니라 특정 조합을 기준으로 하여 가장 적합한 항암 요법을 제시할 수 있을 것으로 기대된다. 이러한 대규모 데이터베이스 구축에는 항암 치료 요법 반응 결과를 포함한 임상 데이터의 공개 및 공유, 데이터 기록 및 관리의 질적 상향 평준화가 도움이 될 것이다.

또한 유전자 변이 검사가 더욱 대중화되고, 검사 비용이 절감될수록 본 도구를 통한 유의미한 조합 발굴에 도움이 될 것으로 기대한다. 본 연구에서 사용된 데이터는 표적 시퀀싱 방식 2건, 전장 엑솜 시퀀싱(WES) 방식 3건으로, 각각 표적 치료 데이터 ①, ②와 면역 치료 데이터①, ②, 간 절제술 데이터이다. 각 데이터에서 추출된 개별 유전자 총 개수는 표적 시퀀싱 데이터의 경우 각각 383개, 163개였으며, WES 시퀀싱 데이터의 경우 562개, 17916개, 15743개였다.

표적 시퀀싱은 특정 유전자 변이만을 표적으로 검사하기 때문에 인식할 수 있는 유전자 수에 한계가 있다. 하지만 WES는 모든 엑솜(exome)을 대상으로 하므로 한 환자당 발견되는 유전자 변이의 개수가 절대적으로 많았다. 실제 추출되는 조합의 수는 환자들의 유전자 변이 양상의 수렴 정도에 따라 조금씩 다를 수 있지만, WES 방식을 사용한 데이터가 표적 시퀀싱을 사용한 데이터에 비해 상대적으로 추출되는 조합 수가 많았다. WES 방식이 더 많은 유전자의 변이를 발견할 수 있음에도 불구하고 질병 표적 유전자 패널을 사용하는 이유는 검사 비용 때문이다. 해를 거듭할수록 유전자 변이 검사 비용이 절감되고 있으므로 WES, 나아가 WGS(Whole Genome Sequencing)의 보편화를 기대할 수 있다[44]. 이에 따라 본 도구를 활용하여 추출될 수 있는 치료 반응과 유의한 상관성을 보이는 유전자 변이 조합의 수와 그 다양성 또한 함께 증가할 수 있을 것으로 기대한다.

2. 기대 효과

가. 표적 치료 요법 임상 데이터 분석을 통한 패스웨이의 확장

본 연구 결과 중 표적 치료 데이터 ①, ②를 통해 추출된 유전자 변이 조합은 본 도구를 활용한 생물학적 패스웨이의 확장 가능성을 보여주었다. 데이터 입력 시 패스웨이에 관련된 사전지식이 전혀 개입되어 있지 않았음에도 유전자 변이 조합과 치료 반응데이터만으로 생물학적 증명이 완료된 유전자 변이 조합을 발견하였다. 이는 결과로 추출된 유전자에 대한 연구가 활발하게 진행되어있기에 증명할 수 있었다. 그러나 암에 작용하는 기전과 그 기능, 상호작용이 밝혀지지 않은 유전자들이 아직도 다수 존재한다. 본 도구를 활용하여 발굴된 유전자 변이 양상을 통해 밝혀지지 않은 생물학적 패스웨이, 단백질 간 상호작용을 발견 및 확장할 수 있을 것이다. 또한 해당 패스웨이에 작용할 수 있는 치료 반응까지 함께 제공함으로써 발견된 패스웨이를 저해할 수 있는 항암 요법 및 항암제를 추천해줄 수 있을 것이다.

나. 면역 치료 요법 임상 데이터 분석을 통한 TMB 관련 유전자 발굴

TMB(Tumor Mutation Burden)란 면역 체크포인트 치료법의 치료 반응을 예측하는 데 있어 가장 잘 알려진 생물지표(biomarker)이며, TMB 값이 큰 암세포가 더 좋은 치료 반응성을 보인다고 알려져 있다. 본 연구는 면역 체크포인트 억제제와 유효한 상관성을 보이는 유전자 변이 패턴을 임상 데이터로부터 추출함으로써 TMB와 직접적, 간접적으로 연관이 있는 유전자 변이를 발굴하였다. 이 중 면역 체크포인트 치료①에서 발굴된 KRAS, TP53 유전자 변이 조합 등은 타 연구 결과로 입증된 조합이었으나, 면역 체크 포인트 치료②의 결과는 잘 알려지지 않은 유전자 변이 조합이 다소 추출되었다. 이 중 일부는 개별 유전자 변이와 면

역 체크포인트 치료와의 연관성을 확인할 수 있었으나, 일부 유전자의 경우 면역 체크포인트 치료, 나아가 암과의 관계성조차 밝혀지지 않은 예도 있었다. 이들은 높은 유의성을 보였음에도 불구하고 해당 유전자에 관련된 기존 연구의 부족으로 설명할 수 없는 한계가 있었다. 본 도구를 활용해 발굴한 면역 치료 요법과 유의한 상관성을 보이는 유전자 변이 조합을 통해 TMB와 관련이 있는 잠재적 유전자 변이를 역추적할 수 있을 것으로 기대한다.

참 고 문 헌

1. Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., … McKinney, E. F. (2019). From big data to precision medicine. Frontiers in Medicine. https://doi.org/10.3389/fmed.2019.00034
2. Le Tourneau, C., Borcoman, E., & Kamal, M. (2019). Molecular profiling in precision medicine oncology. Nature Medicine, 25(5). https://doi.org/10.1038/s41591-019-0442-2
3. “The Cancer Genome Atlas Program - National Cancer Institute”, accessed Jul 01. 2020. URL : https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
4. Tainsky, M. A. (2009). Genomic and proteomic biomarkers for cancer: A multitude of opportunities. Biochimica et Biophysica Acta - Reviews on Cancer. https://doi.org/10.1016/j.bbcan.2009.04.004
5. Goodman, A. M., Kato, S., Bazhenova, L., Patel, S. P., Frampton, G. M., Miller, V., … Kurzrock, R. (2017). Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Molecular Cancer Therapeutics, 16(11). https://doi.org/10.1158/1535-7163.MCT-17-0386
6. Fancello, L., Gandini, S., Pelicci, P. G., & Mazzarella, L. (2019). Tumor mutational burden quantification from targeted gene panels: Major advancements and challenges. Journal for ImmunoTherapy of Cancer. https://doi.org/10.1186/s40425-019-0647-4
7. Mina, M., Raynaud, F., Tavernari, D., Battistello, E., Sungalee, S., Saghafinia, S., … Ciriello, G. (2017). Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies. Cancer Cell, 32(2). https://doi.org/10.1016/j.ccell.2017.06.010

8. Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. Nature. https://doi.org/10.1038/nature10762
9. Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Berghe, W. Vanden, Goethals, B., & Laukens, K. (2015). A primer to frequent itemset mining for bioinformatics. Briefings in Bioinformatics, 16(2). https://doi.org/10.1093/bib/bbt074
10. cBioPortal for Cancer Genomics, accessed Jul 03. 2020. URL : https://www.cbioportal.org/
11. Schwartz, L. H., Litière, S., De Vries, E., Ford, R., Gwyther, S., Mandrekar, S., … Seymour, L. (2016). RECIST 1.1 - Update and clarification: From the RECIST committee. European Journal of Cancer, 62. https://doi.org/10.1016/j.ejca.2016.03.081
12. Hyman, D. M., Piha-Paul, S. A., Won, H., Rodon, J., Saura, C., Shapiro, G. I., … Solit, D. B. (2018). HER kinase inhibition in patients with HER2-and HER3-mutant cancers. Nature. https://doi.org/10.1038/nature25475
13. Razavi, P., Dickler, M. N., Shah, P. D., Toy, W., Brown, D. N., Won, H. H., … Chandarlapaty, S. (2020). Alterations in PTEN and ESR1 promote clinical resistance to alpelisib plus aromatase inhibitors. Nature Cancer. https://doi.org/10.1038/s43018-020-0047-1
14. Sharma, P., & Allison, J. P. (2015). The future of immune checkpoint therapy. Science. https://doi.org/10.1126/science.aaa8172
15. Rizvi, H., Sanchez-Vega, F., La, K., Chatila, W., Jonsson, P., Halpenny, D., … Hellmann, M. D. (2018). Molecular determinants of response to anti-programmed cell death (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell lung cancer profiled with targeted next-generation sequencing. Journal of Clinical Oncology, 36(7), 633 - 641. https://doi.org/10.1200/JCO.2017.75.3384

16. Miao, D., Margolis, C. A., Vokes, N. I., Liu, D., Taylor-Weiner, A., Wankowicz, S. M., … Van Allen, E. M. (2018). Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. Nature Genetics. https://doi.org/10.1038/s41588-018-0200-2
17. Ahn, S. M., Jang, S. J., Shim, J. H., Kim, D., Hong, S. M., Sung, C. O., … Kong, G. (2014). Genomic portrait of resectable hepatocellular carcinomas: Implications of RB1 and FGF19 aberrations for patient stratification. Hepatology. https://doi.org/10.1002/hep.27198
18. den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., Mcgowan-Jordan, J., … Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human Mutation. https://doi.org/10.1002/humu.22981
19. “File Format: MAF - GDC Docs” accessed May 10. 2020. URL : https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
20. “HUGO Gene Nomenclature Committee” accessed Jul 03. 2020. URL : https://www.genenames.org/
21. “OncoTree” accessed Jul 01. 2020. URL : http://oncotree.mskcc.org/#/home
22. Chon, K. W., Hwang, S. H., & Kim, M. S. (2018). GMiner: A fast GPU-based frequent itemset mining method for large-scale data. Information Sciences. https://doi.org/10.1016/j.ins.2018.01.046
23. Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger Koll-Stobbe(eds.) New Approaches to the Study of Linguistic Variability. Language Competence and Language Awareness in Europe, Vol. 4. Frankfurt: Peter Lang. pp. 3-11.
24. Arias-Romero, L. E., Villamar-Cruz, O., Pacheco, A., Kosoff, R., Huang, M., Muthuswamy, S. K., & Chernoff, J. (2010). A Rac-Pak signaling pathway is essential for ErbB2-mediated transformation of human breast

epithelial cancer cells. Oncogene. https://doi.org/10.1038/onc.2010.318

25. He, H., & Huynh, N. (2015). p21-activated kinase family: promising new drug targets. Research and Reports in Biochemistry, 119. https://doi.org/10.2147/rrbc.s57278
26. Wilson, M. R., Reske, J. J., Holladay, J., Wilber, G. E., Rhodes, M., Koeman, J., … Chandler, R. L. (2019). ARID1A and PI3-kinase pathway mutations in the endometrium drive epithelial transdifferentiation and collective invasion. Nature Communications. https://doi.org/10.1038/s41467-019-11403-6
27. Wang, X., Enomoto, A., Asai, N., Kato, T., & Takahashi, M. (2016). Collective invasion of cancer: Perspectives from pathology and development. Pathology International. https://doi.org/10.1111/pin.12391
28. Samartzis, E. P., Gutsche, K., Dedes, K. J., Fink, D., Stucki, M., & Imesch, P. (2014). Loss of ARID1A expression sensitizes cancer cells to PI3K- and AKT-inhibition. Oncotarget. https://doi.org/10.18632/oncotarget.2092
29. Hennessy, B. T., Smith, D. L., Ram, P. T., Lu, Y., & Mills, G. B. (2005). Exploiting the PI3K/AKT pathway for cancer drug discovery. Nature Reviews Drug Discovery. https://doi.org/10.1038/nrd1902
30. Saito, N., Hirai, N., Aoki, K., Suzuki, R., Fujita, S., Nakayama, H., … Iwabuchi, S. (2019). The oncogene addiction switch from NOTCH to PI3K requires simultaneous targeting of NOTCH and PI3K pathway inhibition in glioblastoma. Cancers. https://doi.org/10.3390/cancers11010121
31. “NOTCH1 (OncoKB)”, accessed Jul 3. 2020. URL : https://www.oncokb.org/gene/NOTCH1
32. Henninger, E. E., & Pursell, Z. F. (2014). DNA polymerase ε and its roles in genome stability. IUBMB Life. https://doi.org/10.1002/iub.1276
33. Wang, F., Zhao, Q., Wang, Y. N., Jin, Y., He, M. M., Liu, Z. X., & Xu,

R. H. (2019). Evaluation of POLE and POLD1 Mutations as Biomarkers for Immunotherapy Outcomes Across Multiple Cancer Types. JAMA Oncology. https://doi.org/10.1001/jamaoncol.2019.2963

34. “G1 to S cell cycle control (Homo sapiens) - WikiPathways”, accessed Jun 10. 2020. URL : https://www.wikipathways.org/index.php/Pathway:WP45#nogo2
35. Dong, Z. Y., Zhong, W. Z., Zhang, X. C., Su, J., Xie, Z., Liu, S. Y., … Wu, Y. L. (2017). Potential predictive value of TP53 and KRAS mutation status for response to PD-1 blockade immunotherapy in lung adenocarcinoma. Clinical Cancer Research. https://doi.org/10.1158/1078-0432.CCR-16-2554
36. Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., … Chan, T. A. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. https://doi.org/10.1126/science.aaa1348
37. Pan, H., Chen, L., Xu, Y., Han, W., Lou, F., Fei, W., Liu, S., Jing, Z., & Sui, X. (2016). Autophagy-associated immune responses and cancer immunotherapy. Oncotarget, 7(16), 21235 - 21246. https://doi.org/10.18632/oncotarget.6908
38. Gao, Q., Cui, Y., Shen, Y., Li, Y., Gao, X., Xi, Y., & Wang, T. (2019). Identifying Mutually Exclusive Gene Sets with Prognostic Value and Novel Potential Driver Genes in Patients with Glioblastoma. BioMed Research International. https://doi.org/10.1155/2019/4860367
39. Zhang, W., Hong, R., Xue, L., Ou, Y., Liu, X., Zhao, Z., … Zhan, Q. (2017). Piccolo mediates EGFR signaling and acts as a prognostic biomarker in esophageal squamous cell carcinoma. Oncogene. https://doi.org/10.1038/onc.2017.15
40. Chen, H., Chong, W., Wu, Q., Yao, Y., Mao, M., & Wang, X. (2019).

Association of LRP1B mutation with tumor mutation burden and outcomes in melanoma and non-small cell lung cancer patients treated with immune check-point blockades. Frontiers in Immunology. https://doi.org/10.3389/fimmu.2019.01113

41. Jiang, H., Martin, V., Gomez-Manzano, C., Johnson, D. G., Alonso, M., White, E., … Fueyo, J. (2010). The RB-E2F1 Pathway Regulates Autophagy. Cancer Research. https://doi.org/10.1158/0008-5472.CAN-10-1604
42. Felder, M., Kapur, A., Gonzalez-Bosquet, J., Horibata, S., Heintz, J., Albrecht, R., … Patankar, M. S. (2014). MUC16 (CA125): Tumor biomarker to cancer therapy, a work in progress. Molecular Cancer. https://doi.org/10.1186/1476-4598-13-129
43. Litchfield, K., Summersgill, B., Yost, S., Sultana, R., Labreche, K., Dudakia, D., … Turnbull, C. (2015). Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. Nature Communications. https://doi.org/10.1038/ncomms6973
44. Van Nimwegen, K. J. M., Van Soest, R. A., Veltman, J. A., Nelen, M. R., Van Der Wilt, G. J., Vissers, L. E. L. M., & Grutters, J. P. C. (2016). Is the 1000 genome as near as we think? A cost analysis of next-generation sequencing. Clinical Chemistry. https://doi.org/10.1373/clinchem.2016.258632

Abstract

An analysis tool of co-occurring genetic alteration driven clinical response of anti cancer therapy

Jiwon Son
Healthcare Management and Informatics
The Graduate School
Seoul National University

Introduction : There are both direct and indirect methods of cancer therapy such as targeted therapy, chemotherapy, and immunotherapy and genomic driven precision oncology has been utilized for recommending suitable one for patients. Variants from sequenced samples helps better understanding of cancer states and therefore consists predictive power of clinical benefit. There are some limitations though, such as variable clinical benefits between patients with single identical genomic modifier of response. We implemented co-mutation based analysis to identify significant genomic modifiers that impels clinical benefit.

Methods : The method of Frequent Itemset Mining was used for extracting possible combinations of variants from clinical trials data with sequenced samples. Both groups (clinical benefit and no benefit)

with presence of certain combinations were counted based on provided patient's clinical profile, and ranked by Bayes factor.

Results : This study developed an analysis pipeline for discovering significant genomic modifier or response from cancer trials. For validation, 2 targeted therapies, 2 immunotherapies, and 1 surgical treatment data was deployed. As for highly ranked modifiers we performed literature review for biological and clinical relevance. In general, co-mutations have reached higher predictability in comparison to single variants. This shows both possibility and necessity of co-mutation based approach to identify biomarkers for cancer therapies.

keywords : precision oncology, genetic alteration, anti-cancer therapy, clinical response, frequent itemset mining, cancer genomics

Student Number : 2016-22035